CDA数据分析师系列丛书

中国青年政治学院教改项目支持

数据新闻实战

刘英华 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书紧密围绕数字媒体环境下新闻工作者在数据新闻制作中的实际需求，基于案例全面介绍了数据新闻制作的流程。本书理论和实践结合，内容包括数据新闻的概念和制作流程，公开数据的获取、申请和搜索方法，数据转换和存储方法，“脏数据”的成因及其表现形式，常见的数据清理和分析工具，基于 OpenRefine 环境清理“脏数据”的过程和方法，数据清理原则，数据合理性分析，缺失数据的预测和时间序列预测等。本书同时阐明了数据可视化的概念，详细介绍了 Tableau 制作数据新闻的方法和技巧，最后介绍了其他常用的数据新闻制作工具。

本书通俗易懂、结构严谨、层次清晰、案例丰富，特别适合网络编辑、新媒体记者、大中专院校相关专业师生阅读，有一定工作经验的数据新闻工作者也可以从本书中学习到大量高级实用的功能和技巧。

图书在版编目（CIP）数据

数据新闻实战 / 刘英华著. —北京：电子工业出版社，2016.9
（CDA 数据分析师系列丛书）
ISBN 978-7-121-29738-0

Ⅰ. ①数… Ⅱ. ①刘… Ⅲ. ①数据处理－应用－新闻学 Ⅳ. ①G210.7

中国版本图书馆 CIP 数据核字(2016)第 200127 号

策划编辑：张慧敏
责任编辑：徐津平
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：17.25　字数：450 千字
版　　次：2016 年 9 月第 1 版
印　　次：2016 年 9 月第 1 次印刷
印　　数：4000 册　　定价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

前　言

写作目的

在大数据环境下，数据新闻作为一种新的报道形态受到了读者的认可和追捧。新闻工作者需要全面提升自己的专业技能，其中之一就是具备数据分析和数据呈现的能力。但无论是国内还是国外，大多数新闻工作者缺乏数据方面的知识，因为传统高校缺乏相应的课程，市场上也难以寻觅相应的图书。

现有的数据新闻方面的书籍主要研究的是数据新闻理论、点评数据新闻作品，缺少数据新闻的实战流程。本书正是为学习数据新闻制作的读者准备的，通过阅读本书可以快速获取数据、清理数据、可视化数据，独立完成数据新闻制作的全过程。

本书内容

第 1 章　数据新闻概述。本章阐述数据新闻的概念、数据新闻制作人才的需求、数据新闻技术要求和制作流程，最后展示并点评了近期数据新闻奖的获奖作品。

第 2 章　获取数据。本章讲解获取数据的方法和具体途径，包括政府、国际组织与第三方机构数据的获取，政府信息公开数据的申请，众包搜集数据及搜索引擎的使用，最后讲解数据的存储和综合案例。

第 3 章　清理和分析数据。本章分析“脏数据”的成因及其表现形式，基于 OpenRefine 环境清理“脏数据”，使用 Excel 简单分析数据，阐明数据清理原则和综合案例。

第 4 章　数据质量分析。本章讲解评估数据合理性的外部合理性检查和内部合理性检查，以及游程检验、抽样分析、缺失数据的预测和时间序列预测。

第 5 章　数据分析及可视化工具应用。本章阐明了数据可视化的概念，介绍常见的数据可视化工具。以 Tableau 为例详细讲解了数据可视化的具体方法，包括创建第一个可视化作品、连接数据、数据视图、高级分析、仪表板、故事和发布，最后分析了三个优秀的 Tableau 作品。

第 6 章　其他数据新闻制作工具。本章讲解其他常用的数据新闻制作工具，包括图表绘制工具库 ECharts、标签云、关系图制作工具 PeoplePlotr 和语义万维网服务 Open Calais，最后使用 HTML5 网站制作模板将所有作品整合。

本书内容全面翔实，操作细节清楚，案例典型，方便学习，素材丰富，有利于强化读者操作能力，提高专业技能。本书提供源文件及资料下载，下载地址 http://www.broadview.com.cn/29738。

读前准备

- Windows 操作系统，互联网接入，IE 浏览器、Firefox 浏览器和 Chrome 浏览器。
- 文本编辑器，如 Windows 中的记事本或者 EditPlus。
- 微软 Office 工具包中的 Excel，版本不限。
- 安装 Java 环境，具体参见本书 3.3.1 小节。
- 如果是大中专学生，可以提前申请 Tableau 免费一年使用权。

排版约定

- 菜单项的名称放在【 】中，如单击【分析】|【创建计算字段】选项。
- 代码使用 Courier New 字体并增加阴影，例如：

```
series: [{                          //设置系列列表
    name: '销量',                   //设置图表系列的名称
    type: 'line',                   //设置图表类型是折线图
    data: [5, 20, 36, 10, 10, 20]   //设置系列数据
}]
```

- 使用“+”表示快捷键的组合，如按【Ctrl】+【C】快捷键。
- 没有特殊说明时，单击和双击分别表示鼠标左键单击和双击。

感谢

首先，感谢购买本书的读者。您的阅读是我写作动力的源泉。数据新闻发展较快，真心希望您在阅读本书后提出宝贵的意见，我们可以共同分析探讨问题，为后续图书的撰写提供素材和经验。

其次，感谢我的爱人和父母。在写作最困难的时候，是他们为我鼓劲加油，支持我完成书稿。父母年迈，但很开心地戴着老花镜帮我校稿。

最后，感谢电子工业出版社的张慧敏编辑、杨嘉媛编辑和戴新编辑，她们的严谨细致和辛勤努力保证了本书的顺利出版。

联系作者

如果您对本书有想法和意见，或者想与作者探讨某个问题，请发送电子邮件至 yinghliu@163.com。

刘英华

2016 年 8 月于北京

目　　录

第 1 章

数据新闻概述

- 数据新闻的概念
- 制作数据新闻
- 数据新闻奖（DJA）获奖作品

1.1 数据新闻的概念

数据新闻，也称数据驱动新闻（Data-Driven Journalism）并不是一个新的概念。**数据新闻的雏形**来自于 1821 年 5 月 5 日曼彻斯特卫报（现在的英国卫报*The Guardian*）在其头版新闻“曼彻斯特在校小学生人数及其年平均消费”[1]中使用的数据表，这也是历史上第一份使用头版数据新闻的报纸，如图 1.1 所示。

DAY SCHOOLS.—*Establishmen*	*Boys*	*Girls.*	*Total.*	*Ann. Exp.*	*Remarks.*
Grammar School	155	....	155	1800	..... Endowments.
Blue Coat ditto	80	....	80	2000	Taught, clothed and boarded.
Green Coat ditto	50	....	50	200	Taught and clothed.
Collegiate Church ditto	....	50	50	40	And offertory money: do. do.
Strangeways ditto	10	....	10	100	.....
St. Mary's ditto	12	12	24	40	(Suppose)—Taught and clothed. Funds arising from Sacramental Offerings.
St. John's ditto	9	....	9	40	(Suppose)—Expences raised by voluntary Subscription.
St. Paul's ditto	20	....	20		
Ladies' Jubilee	....	30	30	250	Taught, clothed and boarded. by voluntary Subscription.
Back King-street	21	....	21	40	(Suppose)—Taught and partly clothed. This School is supported by the benevolence of a single individual.
National Schools, Granby-row	194	119	313	600	Voluntary Subscription, and Collections at Churches.
Bolton-street, Salford	300	170	470		
	851	381	1232	£5110	
Dissenters.					
Lancastfrian School, Marshall-st.	692	225	917	400	Voluntary Subscription.
Unitarian, Mosley-street	....	35	35	50	Ditto ditto
Catholic	198	121	319	104	Ditto ditto
SUNDAY SCHOOLS. *Establishment.*	890	381	1271	£554	
Collegiate Church, Shude Hill.	201	205	406		
St. Ann's, Back King-street	50	56	106		
St. Mary's, Back South Parade	130	110	240		
St. Paul's, Green-street	170	183	353		
Turner-street	68	71	139		
Jersey-street	314	281	595		
St. George's, St. George's	141	112	253		
St. John's, St. John's-street	118	163	281		
St. James's, St. James's-street	102	198	300		
St. Michael's, Miller-street	234	352	586	£1023	
St. Peter's, Jackson's-row	..	120	120		
Alport Town	90	....	90		
St. Clement's and St. Luke's, Bennet-street	835	1071	1906	.....	This is, perhaps, the largest School in the Kingdom. It cost about £2,300, of which £512 0 10½ was contributed in small sums by the Teachers and Scholars.
St. Stephen's, Bloom-street	181	297	478		
Oldfield-road	139	204	343		
Trinity, King's Head Yard	220	300	520		
Hulme, Duke-street	185	189	374		
All-Saints, Oxford-road	196	191	387	30	
Ardwick	60	110	170	25	

图 1.1　曼彻斯特卫报头版数据新闻“曼彻斯特在校小学生人数及其年平均消费”部分截图

1　http://www.theguardian.com/news/datablog/2011/sep/26/data-journalism-guardian#data。

这份数据表是对原始数据[1]进行简单数据清理和分析后得到的。

数据新闻的发端，新闻人可以通过计算机辅助分析尝试找出新闻背后的真相。1967 年，美国密歇根州底特律发生严重的黑人骚乱，史称“第十二街骚乱”。在这场骚乱中，支持者及旁观市民与警方发生激烈冲突，最终演变成美国历史上死亡人数最多的骚乱事件之一。当时成千上万的人聚集在底特律街头，造成 43 人死亡、1189 人受伤、7200 人被捕、2000 多栋房屋被毁的惨剧。

惨剧发生后，菲利普·梅耶（Philip Meyer）接受《底特律自由报》（*Detroit Free Press*）的临时安排，采用调查研究和计算机分析[2]，探究编辑和记者们在报道中一直反思的问题——骚乱者究竟是哪些人？骚乱的原因是什么？

在调查前，很多人认为发生骚乱的原因是骚乱者生活贫困、缺乏教育、对没有工作的状况不满。而实际上，菲利普·梅耶对骚乱者所居住的区域随机选择了一些家庭进行调查，收集了 437 人的年龄、就业、收入、种族、受教育程度等基本信息，当然也包括这些人是否上街参加了骚乱。通过对收集的信息进行计算机分析，表明骚乱者既有上过大学的人，也有高中辍学者，因此“受教育程度和收入并不能预测一个人是否会参加骚乱”，参加骚乱的人并不是因为他们学历低或者失业。这项研究也成为历史上采用计算机辅助报道的最早例子之一。

梅耶的报道中没有数据可视化。相反，数据被用来作为证据，证明当时盛行的有关骚乱者的看法是错误的。随着信息时代的到来，越来越多的公共和私有数据被储存和开放，新闻工作者也尝试使用数据来发现问题、解释世界，而如何使得这些复杂、庞大、枯燥的信息变得可靠、透明、通俗易懂，则是对数据新闻工作者最大的挑战。

随着数据新闻的发展，计算机辅助新闻报道的使用得到了巨大发展。伊利诺伊大学调查性报道专业的骑士会会长 Brant Houston 在 1999 年版的尼尔曼报告和著作《计算机辅助报道实用指南》中梳理了近几十年数据新闻的发展。在 20 世纪 70 年代，菲利普·梅耶 （Philip Meyer）继续与地方报纸《费城问询报》（*Philadelphia Inquire*）合作分析了当地司法系统的量刑模式，与他人合作在《迈阿密先驱报》（*Miami Herald*）分析了资产评估记录。菲利普·梅耶 （Philip Meyer）还出版了著作《精确新闻》（*Precision Journalism*），提倡运用社会学、统计学的调查分析方法来报道新闻，遵循定量研究的规范，以达到新闻报道的客观、公证和中立。

20 世纪 80 年代，仅有寥寥可数的几位记者将数据分析融入新闻调查和报道中。

1989 年，《亚特兰大宪法报》（*The Atlanta Journal-Constitution*）的Bill Dedman使用计算机辅助的新闻报道“Investigative Reporting”获得了普利策新闻奖[3]。同一年，Jaspin在密苏里新闻学院建立了美国计算机辅助报道协会（The National Institute for Computer-Assisted Reporting，即NICAR）的前身。1994 年，NICAR正式创办。涵盖互联网基本使用、电子表格和数据库管理等计算机辅助新闻报道的相关课程开始被多个国家的记者接受和学习。

2005 年，美国的开放数据运动正式开始，新闻报道中数据可视化的应用得到了空前的发展，记

1　https://docs.google.com/spreadsheet/ccc?key=0AonYZs4MzlZbdDB0bUl0LWEtczJseGpCRGot。

2　https://en.wikipedia.org/wiki/Philip_Meyer。

3　http://www.pulitzer.org/awards/1989。

者和程序员结合得更加紧密。

当下的数据新闻，深入挖掘数据，可视化呈现数据并合成新闻故事。2011 年 8 月 6 日晚至次日凌晨，伦敦北部托特纳姆地区发生了骚乱，造成至少 8 名警员受伤。2011 年 8 月 7 日晚间，伦敦多个地区发生袭警、抢劫和纵火等案件，警方逮捕了 100 多名肇事者。

在这些骚乱后，首相及其领导的保守派政客把矛头对准了社交媒体，他们一致认为，暴徒通过脸谱网（Facebook）、推特网（Twitter）和黑莓信使（BlackBerry Messenger，BBM）等平台发布煽动性言论，方便组织暴徒并沟通，因此社交媒体是引发这些骚乱的罪魁祸首。

因为英国政府没有对骚乱发生的起因展开调查，所以英国《卫报》与伦敦政治经济学院（The London School of Economics and Political Science）展开合作，尝试用数据分析的方法发现骚乱背后的真相，即发生骚乱的真实原因和后果，谁是趁乱打劫者，他们为何要参与抢劫[1]，以及社交媒体在骚乱中扮演的角色。

“发现骚乱背后的真相”由《卫报》“特别企划”栏目的编辑保罗·路易斯（Paul Lewis）负责，他在骚乱发生期间走遍了全国骚乱的第一现场，并且大部分的报导通过其个人微博账号@paullewis发布。

“发现社交媒体的真相”由英国曼彻斯特大学的罗伯·普克特教授负责。针对推特网提供的 260 万条有关暴动的信息进行数据分析，尝试分析谣言在推特网中的传播模式、不同用户与参与者在信息流的宣传和传播方面所起的作用，以及推特网是否被用于煽动骚乱等[2]。

如图 1.2 所示，使用地图标示出了骚乱发生的地点，并且使用热力图用不同的颜色展示伦敦各地区的经济情况。蓝色代表富裕地区，红色代表贫穷地区（可根据本页脚注 2 的URL查看原始网页，红色在中心区域，蓝色在边角区域）。地图明确标示了骚乱爆发的地点，也可视化地阐明了骚乱发生的原因[3]。

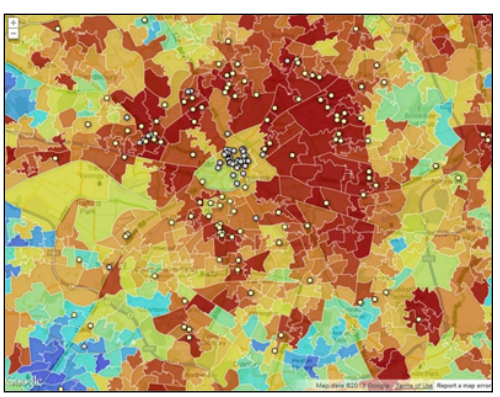


图 1.2　骚乱发生的地点

1　http://www.theguardian.com/uk/series/reading-the-riots。

2　http://www.theguardian.com/news/datablog/interactive/2011/aug/09/uk-riots-incident-map。

3　案例分析来源于英国莱切斯特大学的法利达·维斯（Farida Vis）教授。

如图 1.3 所示的可视化作品[1]呈现了暴动地点与暴徒家庭住址之间的关系，尝试证明骚乱与位置是否存在联系。《卫报》与ITO世界（ITO World）共同模拟出暴徒到达不同地点实施趁火打劫时最有可能经过的路线，突出不同城市的迥异模式，有时候暴徒会长途跋涉到达骚乱地点。

图 1.3　暴动地点与暴徒家庭住址之间的关系

如图 1.4 所示的可视化作品[2]呈现了社交媒体与骚乱的关系，意在说明谣言在推特网上的传播方式。

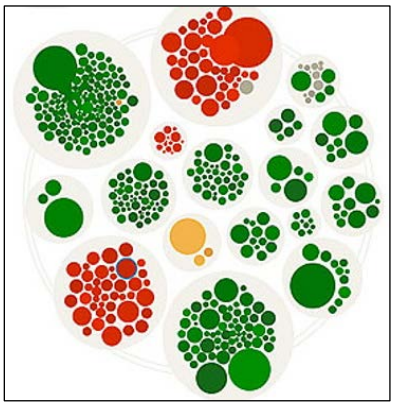


图 1.4　社交媒体与骚乱的关系

罗伯·普克特教授（Professor Rob Procter）领导的学术小组分析了七种谣言，首先收集与每种谣言相关的所有数据并设计出编码表，再根据以下四种主代码对微博信息进行编码：重复谣言者（发表声明）、抗拒者（提出针锋相对的言论）、质疑者（提出疑问）和只评论者（点评）。分析的结果是社交网络只是一种简单的工具，社交媒体用于我们认为正确的活动时，如清理骚乱或阿拉伯之春等，社交媒体的力量似乎是明确积极的，但在其他情况下（未获得我们的认可时），社交媒体往往被描绘

1　http://www.theguardian.com/news/datablog/2011/dec/09/data-journalism-reading-riots。

2　http://www.theguardian.com/news/datablog/2011/dec/09/data-journalism-reading-riots。

成邪恶的、发挥反面作用的。

2013 年 11 月 22 日上午 10 时 25 分，位于青岛经济技术开发区秦皇岛路与斋堂岛街交叉口处的东黄输油管道原油泄漏现场发生爆炸，财新网 [1]记者使用装配有GPS的手机在事发地拍摄了照片，并把这些照片按其所在地点放到了一个互动地图上 [2]，如图 1.5 所示。

图 1.5 “青岛中石化管道爆炸”数据新闻

数据新闻“青岛中石化管道爆炸”可视化呈现了爆炸地点、死难者地点等无法用语言准确表达的信息。该数据新闻的核心工作是将爆炸现场拍摄的照片按拍摄位置映射到谷歌地图，使读者有身临其境的感觉。谷歌地图和照片的结合使读者了解山东、青岛及发生爆炸的黄岛区的具体位置，配合文字描述，将事件发生的时间、地点和起因等进行了完整的描述。

2014 年 6 月，此数据新闻获得亚洲出版业协会（SOPA）的卓越新闻奖，这是中国新闻史上第一次有程序员获得新闻奖。

DDJ（Data-Driven Journalism）出现于 2009 年，中文翻译为数据驱动新闻或数据新闻。2010 年互联网之父——蒂姆·伯纳斯–李（Tim Berners-Lee）宣称数据分析将成为未来新闻的特征。随着 2010 年 10 月 23 日《卫报》刊登的一则“Wikileaks Iraq war logs: every death mapped（维基解密伊拉克战争日志：每一个死亡映射）”[3]（将维基解密提供的数据使用谷歌地图提供的免费软件Google fusion制作了一幅点图，将海湾战争后伊拉克的每一次伤亡事件标示在地图上），将数据新闻带入公众的视野，如图 1.6 所示。

1 http://www.caixin.com/。

2 http://datanews.caixin.com/2013-11-25/100609098.html。

3 http://www.theguardian.com/world/datablog/interactive/2010/oct/23/wikileaks-iraq-deaths-map。

图 1.6　“维基解密伊拉克战争日志：每一个死亡映射”数据新闻

《卫报》的数据新闻记者 Simon Rogers（现加盟推特网数据产品部）认为：“我们可以将深奥难懂的数据做成漂亮、吸引人的样子，并且将这些数据背后的故事讲给想听的人”。

《纽约时报》的阿隆·菲尔霍夫（Aron Pilhofer）认为：“数据新闻是一个概括性术语，它囊括了一套仍在不断增多的用于新闻叙事的工具、技巧与方法，涵盖了从传统的计算机辅助报道（使用数据作为‘信源’）到最前沿的数据可视化和新闻应用等一切叙事方式。其统一的目标是新闻业意义上的‘提供信息和分析以告知我们一天内所有最重要的事件’”。

《芝加哥论坛报》的布莱恩特·博耶（Brain Boyer）认为：“‘数据新闻’和‘文字新闻’的唯一不同在于我们使用了不同的工具包。我们都以探寻、报道和讲述故事为主。‘数据新闻’就像是‘图片新闻’，无非是把相机换成了笔记本电脑”。

万维网创始人蒂姆·伯纳斯–李认为：“数据驱动的新闻代表着未来。新闻工作者需要精通数据。过去你可能通过在酒吧和人聊天获取新闻故事素材，尽管现在这种方式有时可能仍被采用，但目前你同样要钻研数据并借助数据工具分析和筛选令人关注的信息，并对信息加以正确的处理，帮助人们真正看到它反映了什么，在这个国家正在发生什么”。

2010 年 8 月，在阿姆斯特丹举行的首届“国际数据新闻”圆桌会议对“数据新闻”的概念进行了界定：“数据新闻”是一种工作流程，它包括了以下几个方面的内容，通过反复抓取、筛选和重组来深度挖掘数据，聚焦专门信息以过滤数据，可视化地呈现数据并合成新闻故事。

新闻工作者可以通过各个应用终端收集用户的个性化信息，并针对用户的兴趣或选择生成新闻并进行定制化的信息推送，让读者真正拥有“我的新闻”。数据新闻是当下一种新型新闻报道形式，随着数据时代的到来，数据新闻发展迅猛，它的出现在一定程度上改变了传统新闻生产的思路和流

程。伴随着计算机和网络技术的快速发展，大数据成为研究热点，可视化也屡屡被提及。数据新闻的可视化分析和信息可视成为数据新闻的研究热点。

在陈为、沈则潜、陶煜波等编著的《数据可视化》一书中对可视化（Visualization）的解释是：可视化是将数据展现为直观的图形，以帮助理解和记忆。可视化历史久远，广泛应用于地图、统计等领域。

可视化在现代科学中有三个主要分支：科学可视化、信息可视化和可视化分析。

- 科学可视化（Scientific Visualization），主要用于处理科学数据，如地理信息、医疗数据等，以自然科学领域为主。我们日常接触到的地图、气象图、CT 等都属于典型的科学可视化。
- 信息可视化（Information Visualization），主要用于处理抽象的、非结构化的、非几何的抽象数据，如金融交易、社交网络和文本数据。传统的信息可视化起源于统计图形学，又与信息图形、视觉设计等现代技术相关。Excel 中的饼图、柱形图、折线图之类是我们每天都可能接触到的信息可视化作品。
- 可视化分析（Visual Analytics），以可视交互界面为基础进行分析推理，综合图形学、数据挖掘和人机交互等技术。可视化分析是综合性学科，与多个领域相关：在可视化领域，有信息可视化、科学可视化与计算机图形学；与数据分析相关的领域包括信息获取、数据处理和数据挖掘；而在交互方面，则有人机交互、认知科学和感知等学科融合。一个简单的理解就是，看 K 线图分析股价涨跌背后的规律应该是最常见的可视化分析。

黄志敏[1]在《财新经验谈：数据新闻入门》一文中阐明，数据可视化在新闻报道中有三种利用方式：辅助理解、用图表讲故事和数据挖掘。其中，辅助理解类似于插图或配图，是将可视化作为文字报道的辅助手段，这也是常见的方式；用图表讲故事是不借助文字报道，独立用图表展示一个完整的故事，或引导用户接受一个结论，在这个过程中，创作一组图表跟写一篇文章类似，要考虑数据的取舍和讲述的先后次序等，可视化跟写作、拍照、录音、摄像和剪辑一样，都是报道的手段；数据挖掘是通过将数据图形化，展示原本被忽略甚至无法发现的特征。

来自Visual.ly[2]的信息可视图（如图 1.7 所示）用图表的方式展示了什么是数据新闻，其本身也是数据新闻的一个例子。

1.2 制作数据新闻

数据新闻只是新闻报道中的一种形式，弥补传统新闻或叙事性新闻无法呈现的效果。数据新闻采用可视化的方法将单调的数据用一种直观、便于理解和更具说服力的方法呈献给读者。数据，特别是大量的数据比采访几十个对象获取的抽样信息做出的结论更客观、更容易阐明观点。

本节从数据新闻制作的人才需求开始介绍，阐明技术要求和制作流程。

1 黄志敏，财新传媒 CTO，数据可视化实验室负责人。

2 http://visual.ly/what-data-journalism。

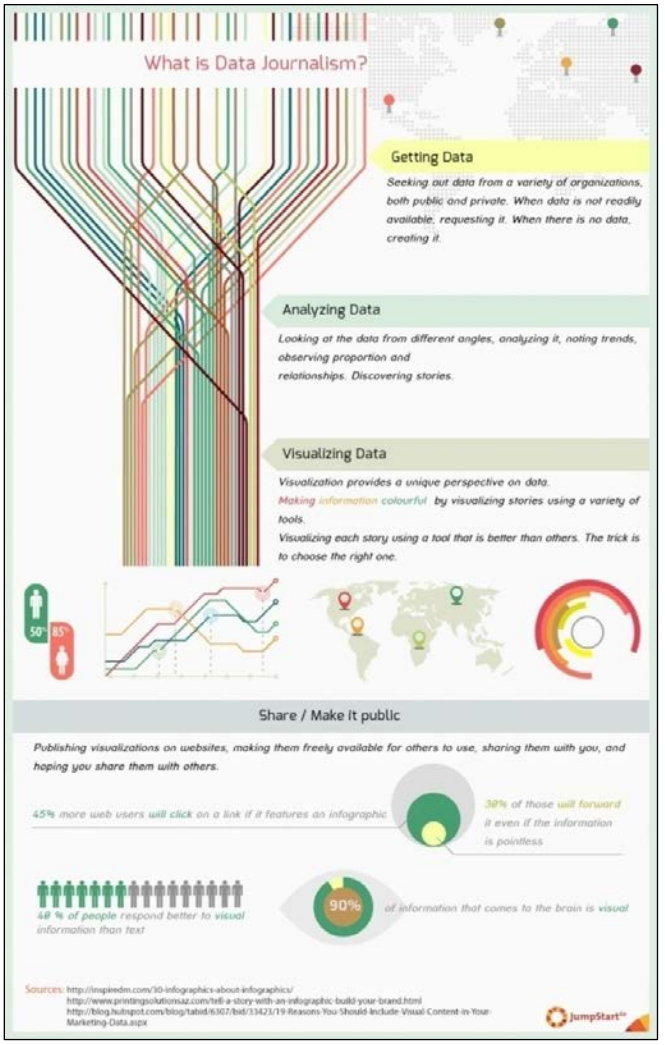


图 1.7　展示什么是数据新闻

1.2.1　人才需求

数据新闻团队一般包含四种角色：记者和编辑、数据分析师、美术设计师和程序设计师。根据数据新闻项目工作量的大小，一个团队可能有 2～3 个人，也可能更多。很多时候一个人需要分饰两个或多个角色，如一个人既是数据分析师，也是程序设计师，或者一个记者，同时也是数据分析师。

记者和编辑的主要工作是采访、写稿、编稿，以及整理相关资源，如与新闻相关的背景资料、图片、视频、音频和文字等。

数据分析师的主要工作是收集和分析数据。具体数据的收集和分析方法参见第 2 章的内容。

美术设计师的主要工作是设计图案，包括手绘图案、3D 制作、图片设计和排版等。使用的主要工具是 Photoshop 和 Illustrator 等。

程序设计师的主要工作是编写代码，实现数据获取和分析、数据可视化等。常见的编程工具有 HTML、Python、R、SQL 和 D3.js 等。

实际上，除了数据新闻团队四个主要角色的工作外，还有视频剪辑、音频剪辑等辅助工作需要完成。

1.2.2 技术需要

最理想的数据新闻人才需要懂新闻、懂数据分析、懂设计且懂编程。而在实际工作中，无论国内还是国外，这样的人才都很难找到，能懂其中两项或三项的复合型人才是现在数据新闻制作团队急需的。本节从技术角度探讨数据新闻制作人才需要掌握的技术。事实上，即使是专业的程序设计师也不可能掌握所有的编程语言和工具，所以本小节将技术分为几类，介绍数据新闻制作需要掌握的技术。

1．数据新闻制作入门级工具

图片、音频和视频编辑工具。数据新闻中往往包含多媒体信息，对新闻图片处理的常用操作包括裁剪照片；加光和减光，即将照片的局部加黑或增亮；修掉照片上由于洗印、扫描、打印而产生的污点；改变照片的反差；对照片的局部进行漂白、清除刮花痕迹等。图片编辑最常用的工具是 Photoshop。对新闻音频的常用操作包括修改采样率、增强与减弱音量、制作淡入和淡出效果、降噪、录音、从视频中提取音频素材、声音特效、声音合成和导出等。音频编辑最常用的工具是 audacity 和 Audition，前者免费而且更容易上手。新闻视频的常用操作包括素材的采集与导入、编辑素材、制作简单特效、添加字幕、混合音频、输出与生成等。视频编辑最常用的工具是 Premiere 和 Final Cut Pro。

数据分析工具 Excel。Excel 是所有数据新闻工作者必须掌握的一个入门级数据分析工具。它用于对数据做简单的清理，如使用函数、分类汇总清理重复记录（案例参见本书 3.6.1 小节“查找重复记录”），使用函数删除多余空格、转换数据类型等，使用分类汇总、排序、数据透视表等完成初步数据分析。

可视化工具 Tableau。Tableau 是一个数据发现、数据分析和数据叙事的数据可视化平台，是数据新闻工作者的入门可视化工具。Tableau 将数据运算与美观的图表完美地结合在一起。它方便地实现了数据连接，无需编程就可以创建地图、条形图、散点图和其他图形，还可以制作数据地图等。

可视化工具 Datawrapper。Datawrapper 是一个在线工具，它可以帮助用户创建交互式数据可视化。这是一个开源工具，能在几分钟内创建可嵌入的图标。因为它是开源的，任何人都可以贡献代码，软件会不断改进。它还包含一个非常棒的图表库，可以查看其他人使用 Datawrapper 完成的作品。

2．数据新闻制作高级工具

爬虫编写工具Python。Python是一种面向对象、语法简洁、大小写敏感的解释型计算机程序设

计语言。它完全免费，简单易学。如果完成同一个任务，假设用C语言要编写 1000 行代码，用Java可能只需要编写 100 行，而用Python可能只需要编写 20 行。读者可以到官方网站[1]下载程序并安装，有很多文档资源也可以在官方网站上找到。制作数据新闻时经常使用Python语言编写爬虫程序，从其他网站抓取数据。对数据新闻工作者来说，学习Python语言的难点是理解正则表达式，可以参考本书 3.3.13 小节“正则表达式”。

数据分析工具 SPSS。SPSS（Statistical Product and Service Solutions）是世界上最早的统计分析软件之一，它是一个专业级的统计分析、数据挖掘、预测分析和决策支持任务的软件产品。数据新闻制作中经常使用 SPSS 实现专业级统计分析和统计图标。有时候获取的数据存在乱码，导入 SPSS 中再导出即可完美解决该问题。例如，本书 2.5.4 小节“结构化信息表格化”中使用 import.io 下载抓取 CSV 的文件有乱码，就可以用上述方法解决。

数据分析工具 R 语言。R 语言是用于统计分析、绘图的语言和操作环境。R 语言属于 GNU 系统，是完全免费而且源代码开放的软件，数据新闻制作时经常使用 R 语言进行统计计算、数据分析和统计制图。

数据可视化工具D3.js。JavaScript是一种直译式脚本语言，而D3.js是一个JavaScript库，可以通过数据来操作文档。D3.js通过使用HTML、SVG和CSS把数据鲜活、形象地展现出来。D3.js严格遵循Web标准，所以其开发的程序兼容主流浏览器。数据新闻制作时经常使用D3.js编写代码，实现在网络上呈现数据的可视化效果，如使用D3.js制作动态图表和漂亮的动态网页地图等。学习D3.js对非IT人士的确是个挑战，但ECharts[2]完美地解决了这个问题。ECharts开源来自百度商业前端数据可视化团队，基于HTML5 Canvas，是一个纯JavaScript图表库，提供直观、生动、可交互、可个性化定制的数据可视化图表。用户可以简单修改代码完成数据可视化，内容详见本书 6.1 节“图表绘制工具库ECharts”。

1.2.3　制作流程

传统新闻制作中更多地体现了记者和编辑、数据分析师、美术设计师和程序设计师的上下游关系，常见的制作流程是：记者首先采访写稿，然后编辑编稿，美术设计师排版配图，最后程序设计师将作品发布到网站上。整个制作过程中美术设计师和程序设计师的参与感相对较差，没有参与开始的选题阶段的工作，导致对作品的了解不是非常全面，理解不到位，从某种角度上来说，可能影响了作品的最后呈现效果。

目前各大媒体也都致力于组建自己的数字新闻团队。每个团队制作数据新闻的具体流程也各有不同，但基本流程是一致的。

在数据新闻制作中，记者和编辑、数据分析师、美术设计师和程序设计师从选题阶段开始就组成了一个团队，共同从各自的专长探讨一个新闻点是否适合做数据新闻、时间或经济成本是否可行。

1　Python 官方网站 https://www.python.org/。

2　ECharts 官方网站 http://echarts.baidu.com。

记者和编辑重点考虑新闻价值。数据分析师重点考虑数据是否可以获取；采用何种工具爬取数据，用 Python 还是 import.io；数据分析采用何种工具；数据分析的时间成本等。美术设计师重点考虑如何手绘图稿及如何排版等。程序设计师重点考虑如何可视化呈现。这种反复的讨论，使团队中的全体成员参与感强，有成就感。

大多数情况下，数据新闻制作时采用“项目”的方式，即一个数据新闻制作团队包含多个项目组，并不是每位成员专属于某一个项目组。很多时候，多个数据新闻项目同时工作，很可能一位成员既属于A项目组，又属于B项目组。目前，我国的数据新闻制作团队往往不会包含太多的成员，以财新数据可视化实验室 [1]为例，团队成员不足 20 人。数据新闻在近几年成为行业的研究热点，单个记者通过再学习，掌握相应的技能后，也可以成为个人数据新闻团队，即通过个人力量收集数据、分析整理数据，可视化呈现，完成数据新闻作品。但从时间成本考虑，多人组成的团队在数据新闻制作中更有竞争力。

数据新闻团队中的四种角色都非常重要，缺一不可。例如，数据新闻的选题也不一定都是记者和编辑提出的，有时候数据分析师在对感兴趣的数据进行分析时，也会发现值得做的新闻点。数据新闻的选题也不一定均出自团队内部，有时候是根据其他记者和编辑提出的外包要求，通过已有的资料实现数据可视化。

数据的获取可能来自于记者和编辑，特别是条线记者，也可能来源于程序设计师编写的代码，如用 Python 编写的爬虫程序，还可能来自于数据分析师的经验（数据分析师更容易了解数据获取的网站），团队里的每个角色都可能从不同的平台、角度获取到合适的数据。

数据分析工作也不一定完全由数据分析师完成，程序设计师、记者和编辑也可能会帮忙。

美术设计工作需要有一定的美学基础，其工作具有一定的特殊性，但记者和编辑更容易从受众的角度给出中肯的设计建议。数据分析师也可以从数据量的角度提出一些设计要求，如图片显示大小等。

数据新闻制作团队的合作精神是非常重要的，现在也提倡在团队中一人分饰多个角色，降低沟通成本。

1.3 数据新闻奖（DJA）获奖作品

由全球编辑网络（Global Editor Network）颁发的数据新闻奖（Data Journalism Awards）一直是数据新闻业界的权威，自 2012 年起每年举办一届。在 2015 年的奖项设置中，除了以往的“年度最佳调查新闻”、“年度最佳新闻应用”等，还加入了“年度最佳数据可视化”、“小型新闻编辑室最佳作品”等 [2]。不同类型的获奖作品，展现了数据新闻的发展，代表了全球最先进数据新闻的水平。

1 财新数据可视化实验室 http://vislab.caixin.com。

2 http://www.globaleditorsnetwork.org/programmes/data-journalism-awards/。

1．年度最佳数据可视化（大型新闻编辑室）

获奖者：美国，华尔街日报，Dov Friedman 和 Tynan DeBold

作品：Battling Infectious Diseases in the 20th Century: The Impact of Vaccines（20 世纪以来和传染病的斗争：疫苗的影响）

网址：http://graphics.wsj.com/infectious-diseases-and-vaccines/

该作品用一条黑线标明了疫苗引进的时间，用不同的色块代表了不同州某一种传染病的病例数量，如图 1.8 所示。通过黑线的左右对比，可以看出在疫苗引进之后，传染病得到了很好的控制。作品容易理解，读者可以快速了解二十世纪以来人类和传染病的斗争历史。

图 1.8　年度最佳数据可视化（大型新闻编辑室）

2．年度最佳数据可视化（小型新闻编辑室）

获奖者：意大利，波森-博尔扎诺自由大学，Matteo Moretti

作品：People's Republic of Bolzano（博尔扎诺人民共和国）

网址：http://www.peoplesrepublicofbolzano.com/

数据新闻“博尔扎诺人民共和国”在 2014 年 9 月以意大利文发表，为了获得更多关注，该专题报道被翻译成英文。

博尔扎诺是意大利的一座城市，当地媒体常常会有类似于“中国人入侵”这样不友好的言论。为了解决公众的负面看法，博尔扎诺自由大学（Free University of Bozen-Bolzano）的 Matteo Moretti 研究员和他的团队组建了一个多媒体网站，向当地人展示他们的担忧毫无根据。

这个多媒体网站包括中国人迁移至博尔扎诺的历史和发展情况、不同年龄段的中国人在博尔扎诺的生活史、博尔扎诺的中国企业和不存在“中国人入侵”四个方面的情况。视频呈现了 8 位身份和特征不同的博尔扎诺中国人的生活故事。通过精准的数据分析和预测中国人在博尔扎诺的所占比

例、分布和人数增量，宏观而直接地证明不存在“中国人入侵”这个结论。

网站使用了 D3.js、Premiere、After Effects 和 Illustrator 等可视化工具，如图 1.9 所示。

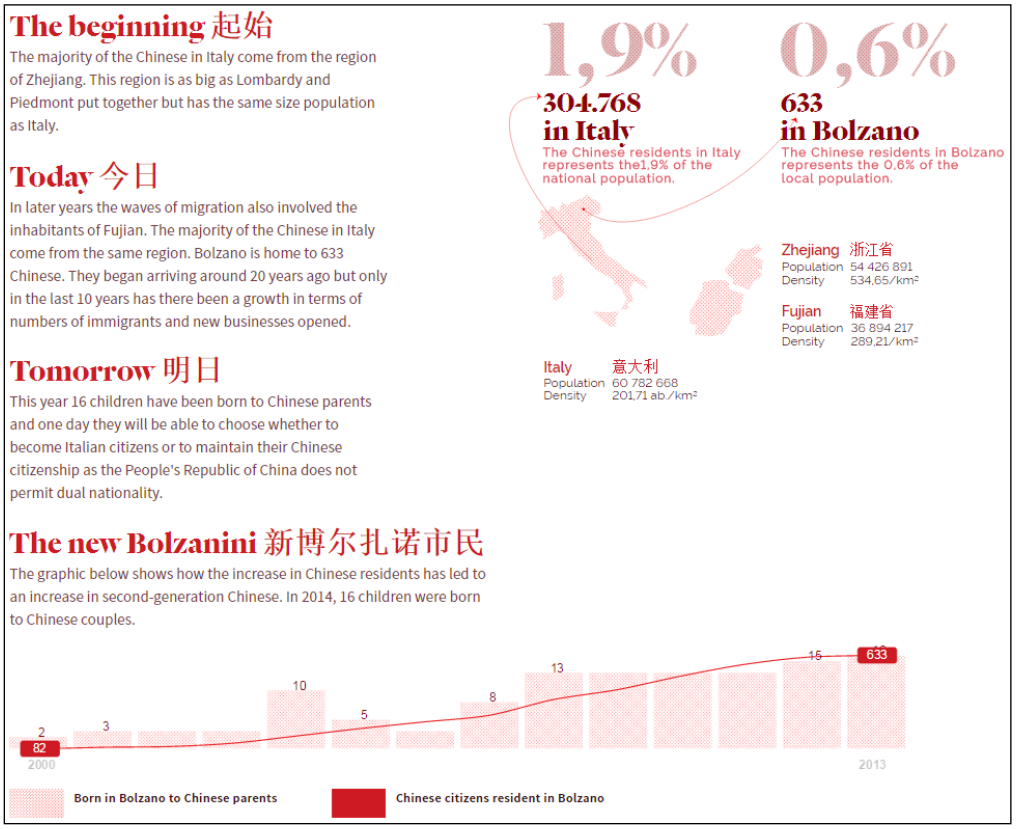


图 1.9　年度最佳数据可视化（小型新闻编辑室）

3．年度最佳调查新闻（大型新闻编辑室）

获奖者：The International Consortium of Investigative Journalists （美国国际调查记者联盟），including Le Monde, The Guardian and 75 media partners both with Mar Cabra - Head of Data and Research Unit（包括《世界报》《卫报》等 75 家媒体合作伙伴和数据研究中心负责人马尔·卡布拉）.

作品：Swiss Leaks - Murky Cash Sheltered by Bank Secrecy and Luxembourg Leaks: Global Companies' Secrets Exposed（瑞士泄密和卢森堡泄密）

网址：http://www.icij.org/project/swiss-leaks

作品由多家媒体合作完成，170 余名记者通过法国的《世界报（Le Monde）》拿到了 60000 份泄露的文件，文件的内容是关于银行如何从遍布世界各地的逃税者和犯罪分子手中获利的。这些泄露的文件包含了超过十万个银行的客户信息。

作品对所有泄露的文件进行重新清理和构建分析，重建了数据库，方便受众理解繁杂的数据，也对银行有很深的警示作用，如图 1.10 所示。

图 1.10　年度最佳调查新闻（大型新闻编辑室）

4．年度最佳调查新闻（小型新闻编辑室）

获奖者：秘鲁，abiola Torres，David Hidalgo，Óscar Castilla，Antonio Cucho 和 Nelly Luna from Ojo Publico

作品：Sworn Accounts: An Analysis of Changes and Wealth of Lima's Mayors（公证账户：利马市长财富的分析）

网址：http://cuentasjuradas.ojo-publico.com/

为了保证秘鲁 10 月份市政和地方选举的公正，来自 Ojo Publico 的五人团队设计了一个数据库检索系统擎，任何人均可通过该系统查看秘鲁首都利马市市长候选人的财务状况。这个数据库包含了 674 位候选人的财务资料，包括跨越 43 个地区的 63 名市长在过去 11 年的数据，数据来自每一个候选人申报的财产及市长的宣誓证词，如图 1.11 所示。

5．年度最佳新闻应用（大型新闻编辑室）

获奖者：英国，BBC News，Bella Hurrell

作品：Which Sport Are You Made For? Take Our 60 Second Test（60 秒测试：你适合哪项体育运动？）

网址：http://www.bbc.com/news/uk-28062001

作品以拉夫堡大学（Loughborough University）运动学专家的一项计算模型为基础建立了一个 APP，进行运动专题的设计，既具有相当高的科学性和权威性，也具有 BBC 的独家发布性。13 个关于身体素质的问题均由弱至强分为十档，在用户进行选择时会自动对问题进行解释，并在选项进度条两侧显示与选项程度相匹配的相应形象化标识，再次帮助用户理解并确认选择无误，简洁巧妙，预计半分钟即可完成的测试时间也符合现代人快速阅读和互动的特征。通过对相关指标的分析，这个应用为用户建立了个人资料，将用户与英联邦运动会中的竞技项目匹配起来，为用户“定制”相

对最合适的三项运动和最不适合的三项运动，并对该运动进行说明，解释测试结果与该运动匹配的原因。

作品利用 R 软件来计算数据和处理信息，再用 D3.js 来设计图形，如图 1.12 所示。

图 1.11　年度最佳调查新闻（小型新闻编辑室）

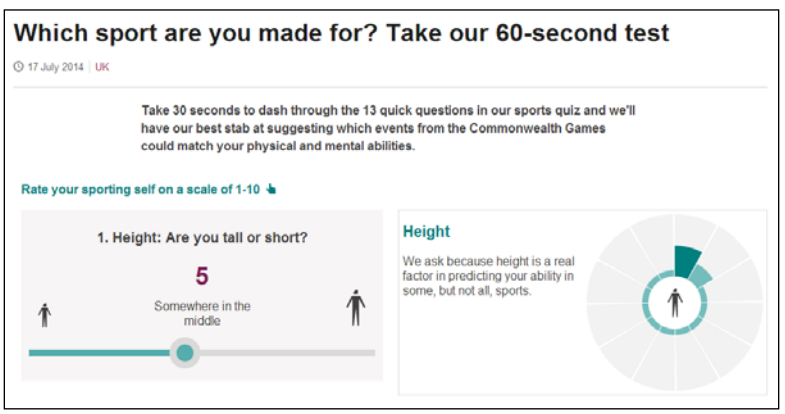


图 1.12　年度最佳新闻应用（大型新闻编辑室）

6．年度最佳新闻应用（小型新闻编辑室）

获奖者：哥斯达黎加，Andrés Fernández A.和 Alejandro Fernández S

作品：Does School Pay Off? How Much?（学校值得你付的学费吗？）

网址：http://www.elfinancierocr.com/gnfactory/especiales/2015/calculadorasalarial

这个 Web APP 的目的是让公民基于哥斯达黎加的可靠数据，采用的科学方法了解教育的回报率（受教育年限和就业市场中的收入间的关系），以激励公民认真学习，防止青少年辍学，如图 1.13 所示。

作品基于 7 个变量（受教育年限、年龄、劳动力市场、劳务分公司、性别、位置和每周工作时间），可以让公民找出就业市场接受的平均月薪，了解 7 个变量的变化带来的工资变化，以及教育年

限与工资的关系。作品设计的模型支持以下趋势：

趋势 1，获取一所中学的毕业证可以增加平均工资的 45%；

趋势 2，一般情况下，获得研究生学位可能会增加一倍的薪水；

趋势 3，受教育年限解释了 20%左右的月工资的差异；

趋势 4，在其他特征不变的情况下，研究生学位获得者在公共部门的就业机会比在私营部门多 62%；

趋势 5，在其他特征不变的情况下，女性工资比男性低 28%；

趋势 6，哥斯达黎加的平均峰值收入年龄为 46 岁。

图 1.13　年度最佳新闻应用（小型新闻编辑室）

7．年度最佳数据新闻网站

获奖者：阿根廷，La Nacion Data team

作品：LA NACION DATA: Open Data Journalism for Change（开放数据新闻，如图 1.14 所示）

网址：http://blogs.lanacion.com.ar/projects/data/la-nacion-data-open-data-journalism-for-change/

8．年度最佳个人作品

获奖者：华尔街日报，数据新闻编辑 Rob Barry

9．年度开放数据奖（两个作品）

获奖者 1：丹麦，《贝林时报（Berlingske）》，Eva Jung 和 Lars Nørgaard Pedersen

作品 1：Tracked（追踪）

网址 1：http://www.b.dk/sporet

获奖者 2：美国，ProPublica 新闻工作室的 Lena Groeger, Charles Ornstein 和 Ryann Grochowski Jones

作品 2：Treatment Tracker: The Doctors and the Services in Medicare Part B（治疗追踪：医疗保险 B 部分的医生和服务）

网址 2：http://projects.propublica.org/treatment/

图 1.14　年度最佳数据新闻网站

获奖作品 1 提供了 2013 年医疗保险为 4900 万老年人和残疾人支付给个人医生和其他保健专业人员的详单，包括各种各样的办公室访问服务、救护车行驶里程、实验室检查、开胸手术医疗费等，还可以使用这个工具来查找和比较供应商，如图 1.15 所示。

图 1.15　年度开放数据奖（作品 1）

获奖作品 2 界面简洁，使用搜索引擎根据输入的州、城市或邮编从大量繁杂的数据中筛选出用户所需内容并可视化呈现，如图 1.16 所示。

图 1.16　年度开放数据奖（作品 2）

10．小型新闻编辑室最佳作品

获奖者：英国，Kiln，Duncan Clark and Robin Houston

作品：The Past, Present and Future of CO2（二氧化碳的过去、现在和未来）

网址：http://www.wri.org/blog/2014/11/past-present-and-future-carbon-emissions

这是 2015 年新设置的一个奖项，可见数据新闻的生产已经普及到小型编辑室，并且受到更多的重视。

该作品展示了 1860 年到 2200 年全球二氧化碳排放的相关情况和问题。不仅使用动态图表和交互式图表可视化呈现了全面的数据，方便读者的阅读和理解，还使用数据进行了预测性报道，如图 1.17 所示。

该作品包含排放（Emissions）、预算（Budget）和未来（Future）三个部分，读者可以自由选择阅读的顺序和方式。排放部分分别呈现了 1860 年至 2013 年碳排放量排名前 20 的国家，数字化直观展示了各国的碳排放情况。预算部分展示了地球平均温度升高 2 度所需的碳排放量，呈现出 1860 年至 2011 年碳排放量的变化及各洲排放比例。未来部分是对历年数据的分析并将计算过程以动态图表的方式呈现，最后还预测性报道油气资源将在 2033 年左右完全耗尽的结论，以及从现在起至 2200 年全球碳排放每年的具体数量和升降情况。

图 1.17　小型新闻编辑室最佳作品

11．特别引荐奖

获奖者：德国，柏林早报（Berliner Morgenpost） 的 Julius Troeger

作品：New and Native Berliners - Who Came, Who Went And Who Lives Here Today（新、老柏林人）

网址：http://www.morgenpost.de/berlin/25-jahre-mauerfall/interaktiv/article136530429/New-Berliners-and-native-Berliners-who-came-who-went-and-who-lives-here-today. html?config=interactive

柏林墙倒塌后的 25 年，大约有 3.5 亿人生活在柏林。随着这个城市人口的增加，住房变得越来越稀缺。仅 2014 年一年，柏林就有约 17 万新居民。柏林墙倒塌后，柏林人口不是连续激增的，而是相互交替的过程，如图 1.18 所示。

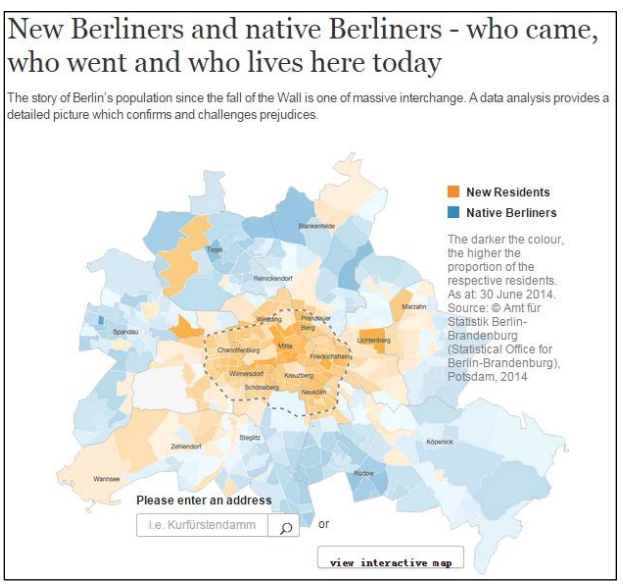


图 1.18　特别引荐奖

12．大众选择奖

获奖者：法国，France Télévisions、Julien Goetz 和 Henri Poulain

作品：Datagueule（数据嘴）（如图 1.19 所示）

网址：https://www.youtube.com/user/datagueule

图 1.19　大众选择奖

第 2 章

获取数据

- 政府、国际组织与第三方机构的公开数据
- 政府信息公开申请数据
- 众包搜集数据
- 搜索引擎的使用
- 数据存储
- 综合案例

数据无处不在，它来自于生活的方方面面。政府收集数据，公司、个人产生数据。随着时间的推移，各行各业产生的数据呈几何递增。数据作为生产要素、无形资产和社会财富，与能源和材料同为重要资源。数据本身最突出的特点是具有重复利用性和增值性，可以在不同的用户中创造不同的价值。

人们往往认为数据的利用和使用需要特别的知识和技能，通常是数学家、工程师或统计分析师才能使用和掌握。随着大数据时代的到来，各行各业都需要与数据打交道，数据在各行各业的影响力与日俱增。特别是在新闻传播行业中，数据可以对社会和环境的状态提供详细而准确的信息，这也正是新闻工作者必须懂得数据并以数据传达信息的原因。数据新闻工作者使用量化技能来收集、分析和传达信息成就了数据新闻的发展和传播。

制作数据新闻最头痛的问题就是“没有数据”。很多时候数据并不是真的没有，而是数据新闻工作者很难在短期内获取所需的数据。例如，有些数据搜集的不完全，或者数据格式不正确，或者多个部门搜集的数据存在矛盾，更多的时候是数据新闻工作者不知道这些数据存储在哪里，以及哪些部门或组织搜集并管理这些数据。

政府作为最大的公共数据资源拥有主体，既要充分利用和共享数据资源，也有责任和义务开放数据资源。很多第三方机构、团体和大学也利用自身的优势收集、整理和开放资源。

2.1 政府、国际组织与第三方机构的公开数据

开放数据资源包含各国政府部门的公开数据资源和非政府机构的数据资源，主要分为免费和收费两种。

1．The ProPublica Data Store（ProPublica 数据库）

网址是 https://projects.propublica.org/data-store/，该数据资源分为以下三种。

- Premium Datasets （收费）

从多种资源搜集、清理并分类的非原始数据集，采取一次性收费政策。

- FOIA Data （免费）

依据美国信息自由法案（FOIA）请求的原始数据是免费提供的，可以自由下载。

- External Data（免费）

仅供在线使用的免费数据集。

2．The Guardian Data Store（《卫报》数据库）

网址是 http://www.theguardian.com/data。

2009 年英国《卫报》开创了“数据博客”，公开了《卫报》数据新闻制作中使用的全部数据，这是数据新闻发展的一个重要里程碑。在“数据博客”页面上，所有数据新闻使用的原始数据均可以免费下载，供读者进一步参考和使用。

3．Google Public Data Explorer（谷歌公共数据资源管理器）

网址是 http://www.google.com/publicdata。

谷歌公司的公开数据库始建于 2010 年，旨在让用户更容易地理解和分享数据。这个在线工具基于著名的 Gapminder Foundation 的 Trendalyzer 软件，主攻时间数据，允许用户创建全面、简洁且互动的可视化图表。

4．World Bank（世界银行数据库）

网址是 http://data.worldbank.org.cn/。

用户可在世界银行数据库中免费获取世界各国的发展数据，该数据库提供超过 9000 个指标文档，可按“国家”下载包含一个国家所有年份的数据，也可按“专题”下载包含所有国家所有年份该专题的指标数据，还可按“指标”下载包含所有国家所有年份该指标的数据。世界银行数据库具有一些高级功能，能够选择和细分数据集、进行定制查询和数据下载、创建图表和其他可视化效果。该数据库还提供包括以表格、地图或图表显示数据的微件服务，微件是一些程序代码，可以嵌入到用户的网页中。

5．UN Data（联合国数据库）

网址是 http://data.un.org/。

联合国数据库为全球用户提供免费数据检索和下载服务。用户可以搜索和下载各种统计资源，包含超过 6000 万个数据点的涵盖范围广泛的主题，如农业、犯罪、教育、就业、能源、环境、卫生、艾滋病毒/艾滋病、人类发展、工业、信息和通信技术、国民账户、人口、难民、旅游、贸易和千年发展目标等。

6．OpenCorporates（OpenCorporates 开放数据库）

网址是 https://opencorporates.com/。

OpenCorporates 是世界上最大的开放式数据库，主要收集公开的公司信息，它提供的各种内部和外部的数据库连接极大地方便了用户。OpenCorporates 还提供公司及管理者网络图，帮助用户了解公司之间的关系及管理者的履历。

7．DataHub（开放知识基金会的数据平台）

网址是 https://datahub.io/。

DataHub 是一个来自开放知识基金会（Open Knowledge Foundation）基于 CKAN 数据管理系统的免费且强大的数据管理平台。

它包含由国家、地方政府、研究机构和其他组织收集的大量数据。凭借其强大的搜索和分类功能，用户可以浏览和找到所需的数据，并可以使用地图、图表和表格等功能。

8．InfoChimps（InfoChimps 数据库）

网址是 http://www.infochimps.com/。

InfoChimps 由数据科学家、云计算和开源专家创立，致力于提供更快、更简捷的大数据系统解决方案。数据库信息包括社交、地理和金融等相关数据。

9．OECD Statistics（经合组织统计信息）

网址是 http://stats.oecd.org/。

OECD 是一个庞大的在线统计数据库，用户可以下载表格，支持多种格式。OECD 对其数据都列出了收集方法和数据源，方便引用和查询。数据集包括 GDP、失业率、教育、金融和医疗等多种类型。

10．NBA 数据

网址是 http://www.basketball-reference.com/。

这个网站统计了 NBA 所有球员、教练、历届比赛的信息和分数，同时也有女篮和奥林匹克赛事等相关数据。

11．美国官方数据库

网址是 http://www.data.gov/。

该网站是美国官方政府的数据库，鼓励公众参与、合作，充分利用联邦政府的数据创建应用、分析产品或科研分析，借此提高政府的透明度和开放度。数据主要来自于大学、联邦政府、州政府和其他非盈利组织等。

12．中华人民共和国国家统计局

网址是 http://data.stats.gov.cn/。

政府是最大的公共数据资源拥有主体，国家统计局网站提供了关于我国土地、水资源、矿产、森林资源、工业状况和人口资源等方面的数据。

13．上海市政府数据服务网

网址是 http://www.datashanghai.gov.cn/。

该网站由上海市人民政府办公厅、上海市经济和信息化委员会牵头，相关政府部门共同参与建设的政府数据服务门户。目标是促进政府数据资源的开发利用，发挥政府数据资源在上海加快建设“四个中心”和具有全球影响力科技创新中心、产业结构调整和经济结构转型中的重要作用，满足公众和企业对政府数据的知情权和使用权，向社会提供政府数据资源的浏览、查询、下载等基本服务，同时汇聚基于政府数据资源开发的应用程序等增值服务。

14．北京市政府数据网

网址是 http://www.bjdata.gov.cn/。

该网站由北京市经济和信息化委员会牵头建设，北京市各政务部门共同参与，于 2012 年 10 月开始试运行。该网站提供北京市政务部门可开放的各类数据的下载与服务，为企业和个人开展政务

信息资源的社会化开发利用提供数据支撑，推动信息增值服务业的发展及相关数据分析与研究工作的开展。

15．Pew Research Center（皮尤研究中心）

网址是 http://www.pewresearch.org

Pew Research Center 是美国的一家独立性民调机构，不间断地发布那些影响美国乃至世界的问题、态度与潮流的信息资料。皮尤研究中心受皮尤慈善信托基金资助，是一个无倾向性的机构。皮尤慈善信托基金资助倡议性项目，包括民意调查、人口统计研究、内容分析和其他数据驱动的社会科学研究等。

16．Dataportals

网址是 http://dataportals.org/。

Dataportals 是一个全面、开放的数据门户网站，包含多个国家、多种语言共 519 个数据门户搜索。

17．美国国会数据库

网址是 http://voteview.com/。

这是乔治亚大学政治学系建立的一个关于美国国会的数据库，包括美国建国至今所有国会议员的投票记录和每个议员的意识形态指数。

18．全球恐怖主义数据库（GTD）

网址是 http://www.start.umd.edu/gtd/。

马里兰大学（University of Maryland）维护的全球恐怖主义数据库是一个开放源代码的数据库，包括 1970 年至今世界各地的恐怖事件的信息（并计划持续更新）。GTD 包括超过 140 000 个国际恐怖事件案例。

2.2 政府信息公开数据的申请

中华人民共和国国务院令第 492 号《中华人民共和国政府信息公开条例》在 2007 年 1 月 17 日国务院第 165 次常务会议通过，自 2008 年 5 月 1 日起施行 [1]。

该条例的实施是数据新闻工作者获取我国各级政府部门在多个领域信息公开的申请基础。虽然各省政府和各级部门申请的流程略有不同，但信息公开申请表的基本信息相同，如图 2.1 和图 2.2 所示为上海市税务系统政府信息公开申请表。如图 2.3 所示为广东省财政厅政府信息公开申请流程图。

该条例公布至今，地方政府部门已经形成了信息公开的习惯，也有一定专业办事的程序，但有些时候一些民间组织或者个人提出的信息公开申请，得到的回复不能令人满意。

1 http://www.gov.cn/zwgk/2007-04/24/content_592937.htm。

例如，2013 年 12 月 2 日，广州市政协委员韩志鹏分别向广东省卫计委、省财政厅和省审计厅快递了《关于公开广东省 2012 年度社会抚养费收支及审计情况的申请》。12 月 4 日，广东省卫计委公布，2012 年度广东省社会抚养费征收总金额是 14.56 亿元。12 月 25 日，广东省财政厅答复，2012 年广东各地征收社会抚养费总额为 26.13 亿元。不同部门的审计结果相差了 11.57 亿元 [1]，如图 2.4 所示。

图 2.1　上海市税务系统政府信息公开申请表——申请人信息 [2]

图 2.2　上海市税务系统政府信息公开申请表——申请信息情况 [3]

1　人民网 http://finance.people.com.cn/n/2013/1226/c1004-23949709.html。

2　https://www.tax.sh.gov.cn/wsbs/WSBS05_zfxxgkSjsq.jsp。

3　https://www.tax.sh.gov.cn/wsbs/WSBS05_zfxxgkSjsq.jsp。

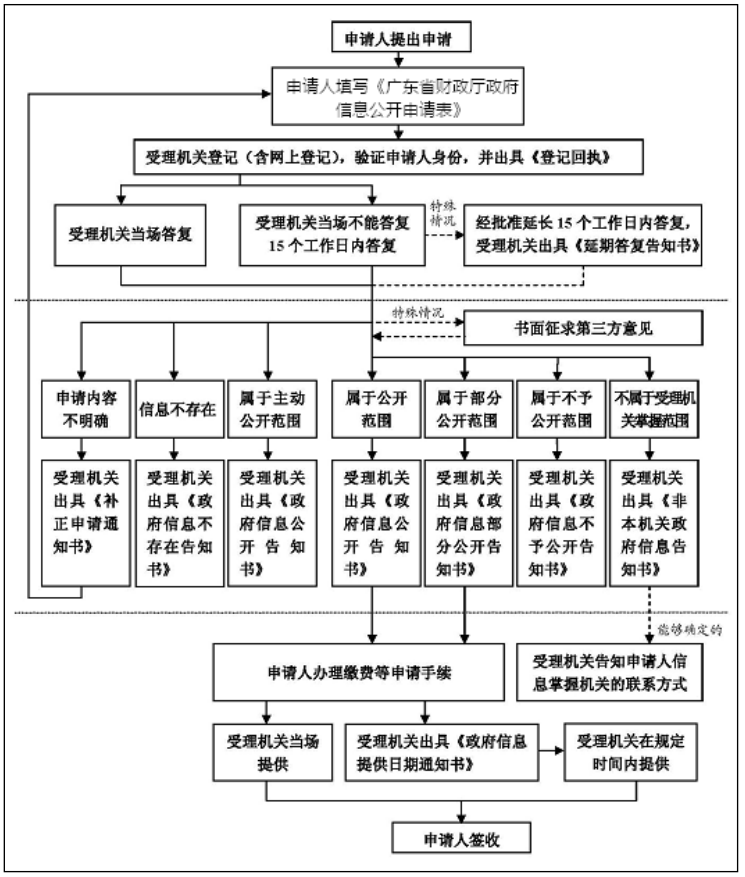
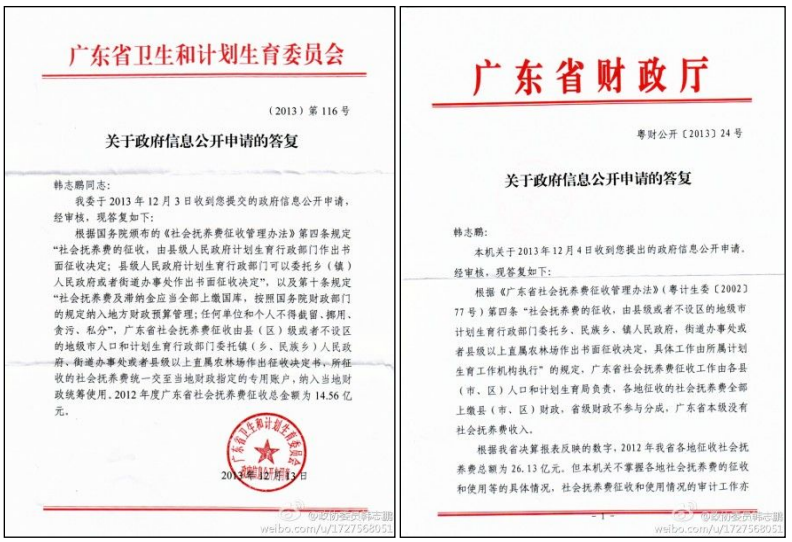


图 2.3　广东省财政厅政府信息公开申请流程图 [1]

例如 2014 年 4 月 4 日，妇女维权志愿者陈思乐发出 320 份收容教育信息公开申请，接收对象是国务院、公安部和 31 个省（区、市）政府和公安厅。具体问题包括“各单位辖区内现有收容教育所的数量与名称，被收教男性与女性人数，被收教者的劳动收入金额和支出去向，收教期限的具体裁量标准，收教期间的收费项目和金额”等，但仅收到了 19 个省公安厅的答复。2014 年 9 月 9 日，陈思乐委托律师向广州市中级人民法院起诉广东省公安厅。这是国内首例关于收容教育信息公开的行政诉讼。2014 年 9 月 18 日广州市中级人民法院正式对此案立案审理 [2]。

1　http://www.gdczt.gov.cn/adminfo/modir/gkzn/201004/t20100425_22159.htm。

2　《凤凰对话 90 后女生：为何起诉粤公安厅》http://news.ifeng.com/a/20140922/42049209_0.shtml。



广东省卫生和计划生育委员会

（2013）第 116 号

关于政府信息公开申请的答复

韩志鹏同志：

我委于 2013 年 12 月 3 日收到您提交的政府信息公开申请，经审核，现答复如下：

根据国务院颁布的《社会抚养费征收管理办法》第四条规定“社会抚养费的征收，由县级人民政府计划生育行政部门作出书面征收决定；县级人民政府计划生育行政部门可以委托乡（镇）人民政府或者街道办事处作出书面征收决定”，以及第十条规定“社会抚养费及滞纳金应当全部上缴国库，按照国务院财政部门的规定纳入地方财政预算管理；任何单位和个人不得截留、挪用、贪污、私分”，广东省社会抚养费征收由县（区）级或者不设区的地级市人口和计划生育行政部门委托镇（乡、民族乡）人民政府、街道办事处或者县级以上直属农林场作出征收决定书，所征收的社会抚养费统一交至当地财政指定的专用账户，纳入当地财政统筹使用。2012 年度广东省社会抚养费征收总金额为 14.56 亿元。

2013 年 12 月 13 日

广东省财政厅

粤财公开〔2013〕24 号

关于政府信息公开申请的答复

韩志鹏：

本机关于 2013 年 12 月 4 日收到您提出的政府信息公开申请，经审核，现答复如下：

根据《广东省社会抚养费征收管理办法》（粤计生委〔2002〕77 号）第四条“社会抚养费的征收，由县级或者不设区的地级市计划生育行政部门委托乡、民族乡、镇人民政府，街道办事处或者县级以上直属农林场作出书面征收决定，具体工作由所属计划生育工作机构执行”的规定，广东省社会抚养费征收工作由各县（市、区）人口和计划生育局负责，各地征收的社会抚养费全部上缴县（市、区）财政，省级财政不参与分成，广东省本级没有社会抚养费收入。

根据我省决算报表反映的数字，2012 年我省各地征收社会抚养费总额为 26.13 亿元，但本机关不掌握各地社会抚养费的征收和使用等的具体情况，社会抚养费征收和使用情况的审计工作亦

图 2.4　两份不同的答复函

2.3 众包搜集数据

如果遍寻网络（包括政府信息公开申请），依旧找不到需要的数据，一般采用调查问卷或者实地调查等方法搜集第一手数据。除此之外，现在还有一种新的数据搜集方法——众包。

《连线》编辑 Jeff Hawe 在 2006 年发表的《众包崛起》中首次提出了“众包”的概念，说明企业利用互联网将工作分配出去的工作方式，以发现创意或解决问题。“众包”应用在数据新闻中就是利用群众的智慧和力量搜集或处理数据，集体完成一个新闻调查计划。最早的“众包”新闻案例是美国《新闻报》的污水系统案例，2009 年《卫报》调查英国议员消费情况无疑是最成功的“众包”新闻案例之一。

2006 年夏天，美国佛罗里达州迈尔斯堡的《新闻报》接到读者举报安装污水系统费用过高的问题。因为需要调查的数据量太大，仅靠专业记者无法在短时间内完成，所以《新闻报》发布了一个举报文件，众多读者根据举报文件展开了调查。最后依托这些读者调查，市政府降低了安装污水系统的费用，还有一名官员因此辞职。通过“众包”调查，《新闻报》在推动事情解决和吸引读者关注两方面都获得了巨大的成功。

2009 年，《每日电讯报》的多篇报道揭露了英国议会议员违规消费的情况。为了回应公众的不满情绪，英国政府在网络上公布了所有议员 4 年来的花费情况，总计约 100 余万份未经整理的原始数据文件。《卫报》设计了一个类似于游戏网站的 Web 页面，邀请读者参与调查议员的花费情况。在调查项目上线的 80 小时内，就有 170 000 份文件被读者审查完毕。

2014 年春节前，《南方周末》联合环保组织“创绿中心”和IT工程师环保公益协会发起了“回乡

测水”行动[1]。“创绿中心”提供低成本、便携、快速、可订制的水质检测工具，让公众有能力、有渠道参与水质检测，同时结合WEB-GIS构建水质信息平台，公众能实时上传水质检测信息，与他人分享。《南方周末》记者汪韬基于此次“众包”调查推出新闻《“回乡测水” 家乡水，清几许？》[2]，IT工程师环保公益协会完成了后期数据可视化[3]。

2.4 搜索引擎的使用

搜索引擎（Search Engine）泛指在网络上以一定的策略搜集信息，对信息进行组织和处理，并为用户提供信息检索服务的工具或系统。搜索引擎被业界公认为继广告、网络游戏和无线增值之后互联网的第四桶金。

搜索引擎工作过程包含三个步骤。首先抓取网页，每个独立的搜索引擎都有自己的网页抓取爬虫（Spider），爬虫顺着网页中的超链接，从一个网站“爬”到另一个网站，通过超链接分析连续访问以抓取更多网页。其次处理网页，搜索引擎“抓”到网页后，要提取关键词，建立索引库和索引。最后，提供检索服务，用户输入关键词进行检索，搜索引擎从索引数据库中找到匹配该关键词的网页。

数据新闻工作者不生产数据，而是搜索和使用数据。搜索指令的合理使用可以帮助用户获取更精准的数据。

2.4.1 搜索指令

使用搜索指令可以帮助用户精准、快速地找到所需的信息。不同的浏览器支持的搜索指令不同，本小节以国内“百度”搜索引擎为例讲解常用的搜索指令。

1．intitle 和 allintitle

intitle 指令将搜索范围限制在网页的标题。allintitle 指令搜索的所有关键字都必须在网页的标题中。如输入“intitle:巴黎恐怖袭击”共搜索到约 193 000 个结果，如图 2.5 所示。输入“allintitle: 巴黎恐怖袭击”共搜索到约 1060 个结果，如图 2.6 所示。

2．intext 和 allintext

intext 指令将搜索范围限制在网页的正文（忽略超链接文本、URL 和标题等）。allintext 指令搜索的所有关键字都必须在网页的正文中。

1 http://tools.ngo20.org/index.php/post/211。

2 http://www.infzm.com/content/98057。

3 http://water.epmap.org/ngo。

图 2.5　使用 intitle 搜索指令

图 2.6　使用 allintitle 搜索指令

3．inurl 和 allinurl

inurl 指令将搜索结果限制在特定 URL 或者网站页面上。allinurl 指令搜索的所有关键字都限制在 URL 或者网站页面上。如仅在政府网站中搜索“巴黎恐怖袭击”，则输入“inurl:gov.cn　巴黎恐怖袭击”，共搜索到约 561 个结果，如图 2.7 所示。如仅在 URL“news.ifeng.com/world”搜索“巴黎恐怖袭击”，则输入“allinurl:news.ifeng.com/world 巴黎恐怖袭击”，共搜索到约 92 个结果，如图 2.8 所示。

图 2.7 使用 inurl 搜索指令

图 2.8 使用 allinurl 搜索指令

4．site

site 指令将搜索限制在站点或者顶层域名上。如仅在特定网站“www.ifeng.com”搜索“巴黎恐怖袭击”，则搜索指令是“巴黎恐怖袭击 site: www.ifeng.com”，如图 2.9 所示。注意，在“site”指令后的站点或顶层域名前不能加“ http://”，如搜索指令“巴黎恐怖袭击 site: http://www.ifeng.com”无法正确执行，如图 2.10 所示。

图 2.9　site 搜索指令正确用法

图 2.10　site 搜索指令错误用法

5．filetype

filetype 指令将搜索限制为某类特定后缀或者文件名的扩展名。如仅搜索“ppt”扩展名的文档，则搜索指令是“巴黎恐怖袭击 filetype:ppt”，如图 2.11 所示，搜索结果均是扩展名为 PPT 的 PowerPoint 文件。

图 2.11　使用 filetype 指令

6．排除（“–”）

“–”代表不包含减号后边的词的页面。使用这个指令时，减号前面必须是空格，减号后面没有空格，紧跟着需要排除的词。如图 2.12 所示，搜索结果是包含“恐怖袭击”但不包含“巴黎”的页面。

7．完全匹配（“”）

完全匹配搜索，即搜索结果包含双引号中出现的所有词，连顺序也必须匹配。例如，搜索“巴黎恐怖袭击”时加双引号，则百度搜索结果完全匹配“巴黎恐怖袭击”。如图 2.13 和图 2.14 所示，对比不使用完全匹配和使用完全匹配搜索结果的不同，使用完全匹配的搜索结果更精准。在当今的大数据时代，使用搜索引擎可以快速搜索到大量的结果，但用户往往没有足够的时间查看数以万条甚至更多的搜索结果，所以精准搜索数据是用户更关注的。

图 2.12　使用排除（“–”）搜索指令

图 2.13　不使用完整匹配搜索指令

图 2.14　使用完整匹配搜索指令

2.4.2　百度搜索工具

百度搜索工具以图形化界面完成搜索指令，如图 2.15 所示。“时间不限”选项可以设置搜索时间条件，如“一月内”或自定义从“2015-12-31”至“2016-6-30”。如图 2.16 所示设置的时间是从“2015-12-31 至今”。“所有网页和文件”选项可以设置搜索到的文档类型，如图 2.16 所示设置的是“PDF 文件”。

“站点内检索”选项可以限制在某个站点或者顶层域名内搜索，如图 2.16 所示设置的是“wenku.baidu.com”。

图 2.15　百度搜索工具界面

图 2.16　使用百度搜索工具

2.4.3　百度高级搜索页面

百度高级搜索页面的网址是 http://www.baidu.com/gaoji/advanced.html，如图 2.17 所示。

百度高级搜索页面可以限定包括或不包括关键字、限定搜索结果显示的条数、限定搜索时间、限定搜索的网页语言、限定文档格式、限定关键词位置和限定搜索位置等。实际上，百度高级搜索集成了常见的搜索指令，用户无需记住复杂的搜索指令就可在图形化搜索界面完成复杂的搜索任务。

图 2.17　百度高级搜索页面

2.5　数据存储

很多时候，通过搜索引擎搜索到的数据并不能获取或分析，如文档是 PDF 格式的、文档是图片

格式的或者 Web 页面的数据是动态的无法抓取等。本节通过一些案例说明如何将其他格式存储为便于分析的 Excel 格式，以及如何获取动态页面数据和批量下载等。

2.5.1　PDF 格式转换为 Excel 格式

许多机构公开发布数据时，都会选择以 PDF 文档呈现，以确保文档内容与格式在不同的设备和平台均可以完美再现，避免内容缺失和格式错位等问题。但 PDF 文档对数据分析的支持太差，一般选择将其转换为便于分析的数据格式，如 Excel 软件的 XLS 或 XLSX 格式。

随着媒体的关注，PM2.5 污染引起了人们的广泛关注。中国环境监测总站[1]作为全国环境监测的技术中心，受到了用户的青睐。该网站每季度发布当季度各月 74 座城市的空气质量状况报告。如“2015 年第三季度 74 城市空气质量状况报告”，Web页面以图片形式呈现，页面底部以PDF格式的附件保存[2]，如图 2.18 所示为“附表 6”。为了更好地分析和梳理数据，该文档需要转换为Excel格式。

附表 6　2015 年第三季度 74 城市 CO-95per 浓度排名情况

单位：mg/m³

排名	城市	CO-95per	排名	城市	CO-95per
1	海口	0.7	18	南京	1.0
1	丽水	0.7	39	淮安	1.1
1	拉萨	0.7	39	江门	1.1
1	厦门	0.7	39	连云港	1.1
5	舟山	0.8	39	合肥	1.1
5	张家口	0.8	39	银川	1.1
5	福州	0.8	39	常州	1.1
5	贵阳	0.8	39	武汉	1.1

图 2.18　“2015 年第三季度 74 城市空气质量状况报告”之“附表 6”

用鼠标右键单击该 Web 页面，在弹出的快捷菜单中选择【复制图片】或【图片另存为】，将页面保存为图片形式。图片中包含数据，但数据无法复制到 Excel，OCR 图像识别软件可以解决这样的问题，但一般需要付费。例如，ABBYY FineReader Professional 是一款真正的专业级 OCR 软件，它既支持多国文字，还支持彩色文件识别、自动保留原稿插图和排版格式及后台批处理识别等功能，主要用于识别扫描图像、图片型 PDF 并转换成可编辑的文本。

选择该 Web 页面底部的附件“附件1：2015年第三季度74城市空气质量报告”并下载 PDF 格式的文件。

如果直接将该文档的“附件 1”～“附件 7”共 7 个表格直接复制、粘贴到 Excel 文件中，则表格的每行数据直接复制到 Excel 的一个单元格中，如图 2.19 所示。

1　http://www.cnemc.cn。

2　http://www.cnemc.cn/publish/totalWebSite/news/news_46447.html。

	A
1	排名 城市 CO-95per 排名 城市 CO-95per
2	1 海口 0.7 18 南京 1.0
3	1 丽水 0.7 39 淮安 1.1
4	1 拉萨 0.7 39 江门 1.1
5	1 厦门 0.7 39 连云港 1.1
6	5 舟山 0.8 39 合肥 1.1
7	5 张家口 0.8 39 银川 1.1
8	5 福州 0.8 39 常州 1.1
9	5 贵阳 0.8 39 武汉 1.1
10	9 台州 0.9 39 苏州 1.1

图 2.19　直接复制、粘贴 PDF 表格到 Excel

使用 Excel 提供的“分列”功能可以实现数据拆分。单击【数据】|【分列】选项，在打开的“文本分列向导”对话框中进行分隔符号的选择，在本例中选择“空格”即可，如图 2.20 所示。分列后的数据如图 2.21 所示，分列后的数据方便后期的数据清理和分析。

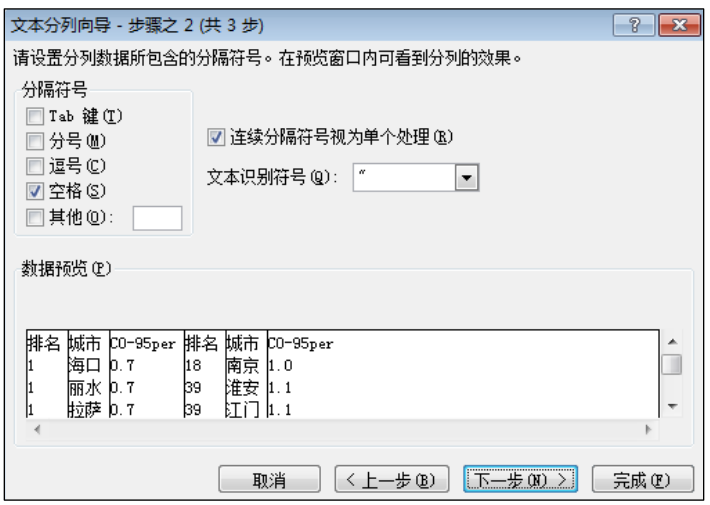


图 2.20　Excel 的“文本分列向导”对话框

	A	B	C	D	E	F
1	排名	城市	CO-95per	排名	城市	CO-95per
2	1	海口	0.7	18	南京	1
3	1	丽水	0.7	39	淮安	1.1
4	1	拉萨	0.7	39	江门	1.1
5	1	厦门	0.7	39	连云港	1.1
6	5	舟山	0.8	39	合肥	1.1
7	5	张家口	0.8	39	银川	1.1
8	5	福州	0.8	39	常州	1.1
9	5	贵阳	0.8	39	武汉	1.1
10	9	台州	0.9	39	苏州	1.1

图 2.21　Excel 分列后的效果

还可以通过PDF转换工具完成转换，如CometDocs[1]、PDF to Excel Online[2]和PDF to Excel[3] 等。

使用 CometDocs 将 PDF 格式的“2015 年第三季度 74 城市空气质量状况报告”转换为 XLSX 格式，如图 2.22 所示。转换后的文件可以下载到本地计算机，如图 2.23 所示。

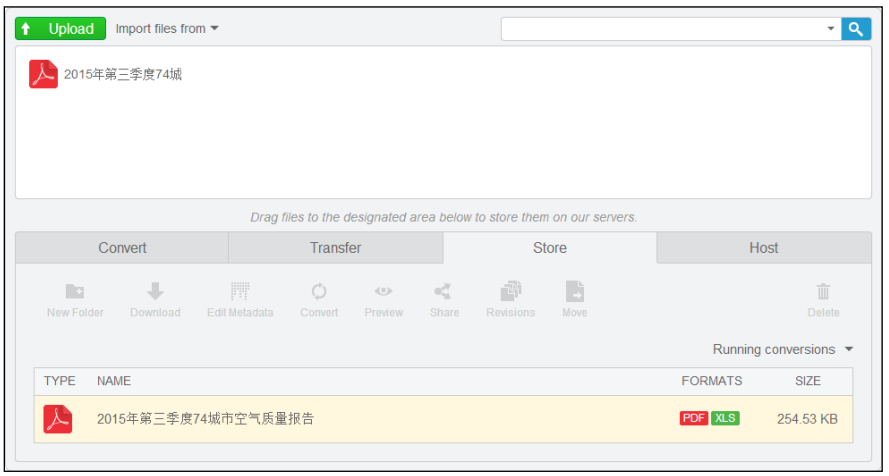


图 2.22　使用 CometDocs 转换 PDF 为 XLSX 格式

图 2.23　下载转换后的 XLSX 文件

2.5.2　在线转换工具 Zamzar

Zamzar 是一个强大且免费的在线转换工具，支持 1200 余种格式转换，包括图片格式、文档格式、音频格式和视频格式等，是一个比较全能的转换工具，而且页面简洁易用，速度快，最重要的

1　http://www.cometdocs.com/。

2　https://www.pdftoexcelonline.com/。

3　http://www.pdftoexcel.org/。

是不需要注册即可使用。网址是 http://www.zamzar.com。

使用 Zamzar 转换文件需要五个步骤，前四个步骤如图 2.24 所示。

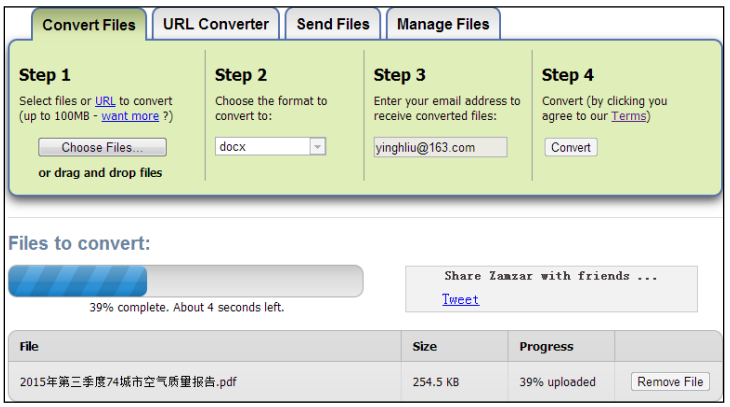


图 2.24　使用 Zamzar 转换文件

第一步，上传文件。单击“Choose Files”按钮选择要转换的文件。Zamzar 允许同时上传一个或多个文件。也可以单击“URL”输入要转换文件的网络地址。免费的 Zamzar 要求转换的文件不能超过 100MB。如果文件过大，可以单击“want more”链接选择付费服务，转换 200MB 以下的文件每月 9 美元，转换 400MB 以下的文件每月 16 美元，转换 2GB 以下的文件每月 49 美元。

第二步，单击下拉按钮选择要转换的格式，如选择“docx”。

第三步，输入接受转换文件的邮箱。

第四步，单击“Convert”按钮开始转换。此步骤显示文件转换的百分比，转换完成后有提示，可以进入邮箱下载。

第五步，进入邮箱，找到 Zamzar 发送的邮件，在文件中找到下载链接地址，单击链接地址进入下载页面即可下载转换后的文件，如图 2.25 所示。

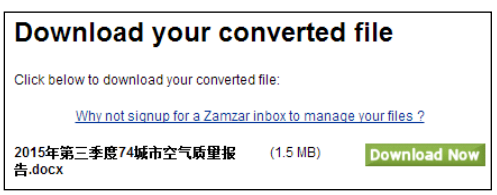


图 2.25　下载 Zamzar 转换后的文件

2.5.3　浏览器插件

中国环境监测总站[1]首页右下方的“空气质量日报”详情是一个**实时更新**的表格，每页显示 9 个

1　http://www.cnemc.cn。

市的基本信息，包含地区、首要污染物、等级和AQI，如图 2.26 所示。其中，AQI是空气质量指数（Air Quality Index）的简称。

如果使用直接复制、粘贴的方法，工作量非常大，而且数据是动态的，很难保证正确抓取全部数据。

使用 Firefox （火狐） 浏览器的 Dafizilla Table2Clipboard 插件，可以完美解决这个问题。

首先需要下载并安装火狐浏览器及插件。登录火狐浏览器的官方网站[1]，选择操作系统后（火狐浏览器支持Windows、Mac OS X、Linux、Linux 64-bit和Android等操作系统）下载最新的版本。正确安装后，单击最右侧的“打开菜单”按钮，在菜单中选择“附加组件”，如图 2.27 所示。

空气质量日报 详情　2015-12-16

地区	首要污染物	等级	AQI
营口市		优	46
阜新市		优	31
辽阳市		优	41
盘锦市		优	49
铁岭市	PM2.5	良	58
朝阳市		优	36
葫芦岛市	PM10	良	54
长春市	PM2.5	良	52
吉林市	PM10,PM2.5	良	55

图 2.26　中国环境监测总站首页的“空气质量日报”

图 2.27　火狐浏览器的“打开菜单”

单击“扩展”选项查看已经安装的插件，若“Table2Clipboard”插件没有安装，可以在“获取附加组件”中下载并安装，如图 2.28 所示。

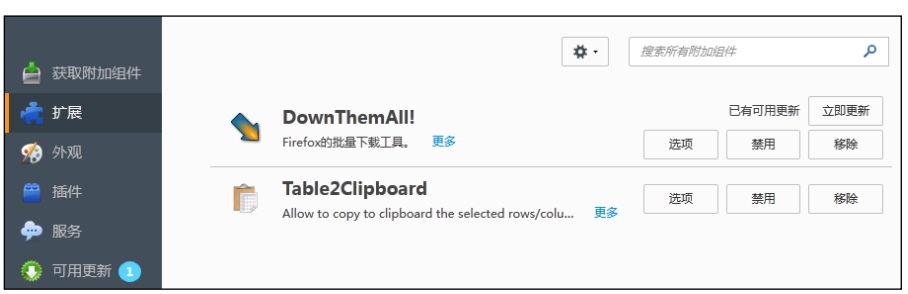


图 2.28　火狐浏览器已安装的插件

使用火狐浏览器打开中国环境监测总站页面，用鼠标右键单击“空气质量日报”中的数据部分，在打开的快捷菜单中选择【Table2Clipboard】|【Copy whole table】选项，如图 2.29 所示。最后打开 Excel，按【Ctrl】+【V】快捷键粘贴即可。

1　http://www.firefox.com.cn/download/。

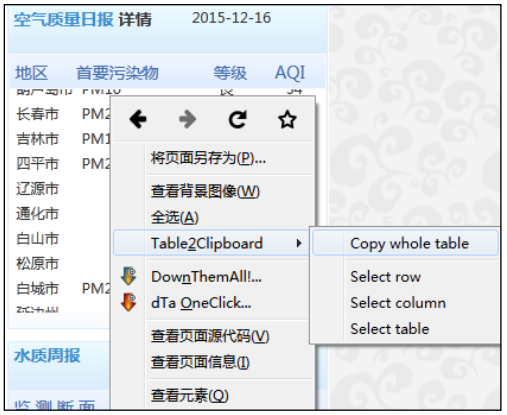


图 2.29　快捷菜单

思考：

登录北京市环境保护监测中心网站 http://www.bjmemc.com.cn，下载当天各监测子站的名称、空气质量指数和空气质量级别。

2.5.4　结构化信息表格化

import.io 是最好的数据提取工具之一，界面非常简单易用，不要求使用者写任何代码即可自动识别网页结构，抓取内容并生成表格供使用者下载。特别适合于抓取内容多且格式统一的 Web 页面。

进入宜家中文主页 [1]，在搜索框中输入“椅子”，单击“搜索”按钮查看页面 [2]。页面上的数据非常规整，图片、相应文字解释和链接排列整齐。这个页面不是以表格形式呈现的，而是做成了图文列表，但呈现出明显的结构化信息，如图 2.30 所示。

该页面的信息通过复制和粘贴操作，无法保存成一个清楚的表格。如果会写代码，可以编写抓取程序自动抓取不同层级的页面资料。但如果不会写代码，则需要通过一些现有的工具，例如 import.io 去抓取。

首先登录import.io网站 [3]，然后申请账号并登录。在网站首页输入图 2.30 对应的URL地址“http://www.ikea.com/cn/zh/search/?query=+%E6%A4%85%E5%AD%90”，单击“Try it Out”按钮，抓取的数据结果如图 2.31 所示。

1　http://www.ikea.com/cn/zh/。

2　http://www.ikea.com/cn/zh/search/?query=+%E6%A4%85%E5%AD%90。

3　https://import.io。

图 2.30　搜索宜家“椅子”页面

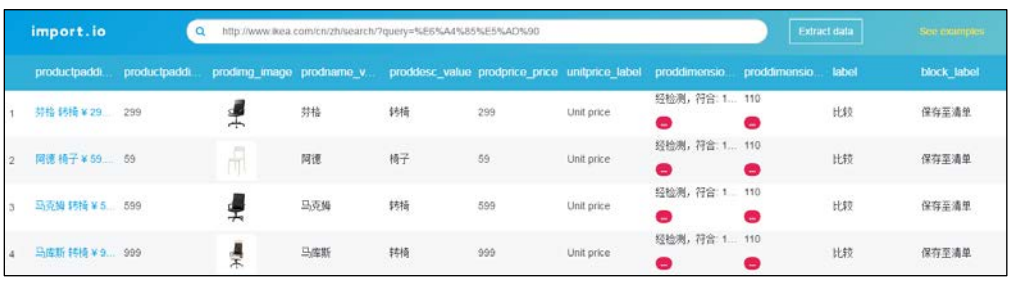


图 2.31　import.io 抓取的数据结果

单击数据结果页面右下角的“Download CSV”按钮，下载抓取的文件，如图 2.32 所示，下载 20 页，然后设置保存的文件名和位置，如图 2.33 所示。

图 2.32　下载页面

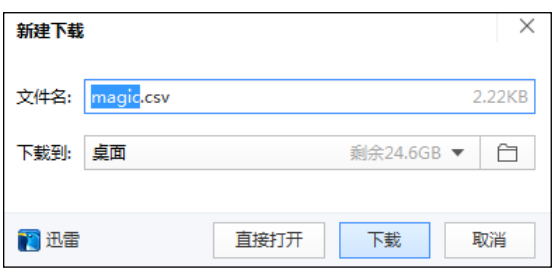


图 2.33　设置文件名和位置

成功案例：What Music Matters Most to KEXP[1]

美国西雅图的地区广播电台KEXP[1]的官方网站上，有一个实时更新的播放列表，将电台所有播

1　http://www.jewelloree.com/2014/03/24/pop-viz-what-music-matters-most-to-kexp/。

放过的音乐都记录下来。数据分析家及音乐爱好者Jewel Loree统计了2013年KEXP电台播放过的所有音乐。她首先使用import.io抓取网站的数据，然后利用Tableau制作了可视化图表，并从不同角度分析数据，例如统计不同时期电台的音乐总播放率。

如图2.34所示是该作品的一部分，是按星期统计的电台音乐播放率，可以看出个别时间的音乐播放次数特别少，如2013年2月23日和2013年9月14日。

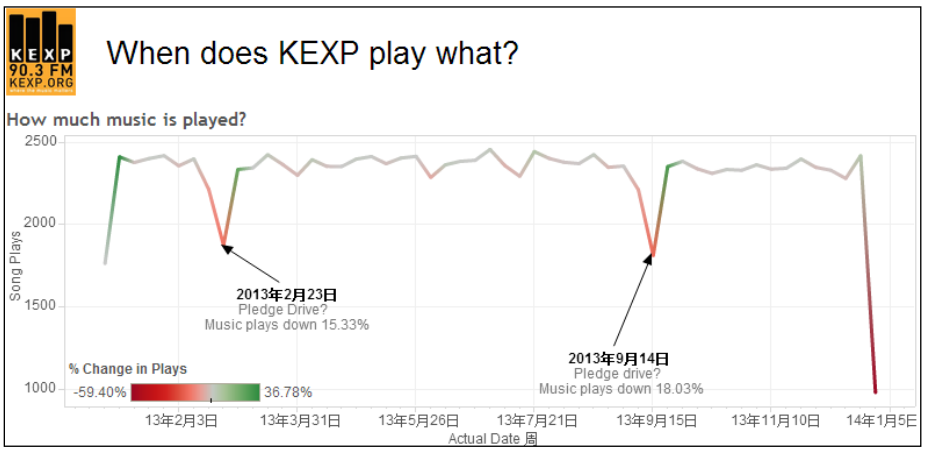


图2.34　作品“What Music Matters Most to KEXP”的一部分

思考：

打开亚马逊首页（http://www.amazon.cn），商品资料排放得整整齐齐，虽然适合展示，却不适用于数据分析，尝试使用import.io将内容转换成表格。

使用kimonolabs[2]也能实现类似的功能，尝试学习该网站的用法。

2.5.5　批量下载文件

有些网页上包含多个链接提供下载，如果手动逐个下载非常浪费时间。

登录北京市环境保护监测中心网站http://www.bjmemc.com.cn，选择“在线服务”导航栏下的“资料下载”[3]，如图2.35所示。可以单独下载每个文件，也可以使用工具一次性下载。常见的解决方法有两种，一种是使用火狐浏览器插件“DownThemAll”下载，另外一种是使用迅雷下载。

1　http://www.kexp.org/。

2　https://www.kimonolabs.com/。

3　http://www.bjmemc.com.cn/g342.aspx。

图 2.35　北京市环境保护监测中心网站

首先在火狐浏览器中安装“DownThemAll”插件（插件安装方法参见 2.5.3 小节），登录北京市环境保护监测中心网站后用鼠标右键单击该页面，然后在打开的快捷菜单中选择【DownThemAll】，在打开的窗口中选择下载的文件类型，本例中勾选“文档”（仅下载 PDF、XLS、DOC 等文档）复选框，在窗口左下角可以看到“选中链接：5/35”，表示当前页面共 35 个链接，过滤器过滤后仅下载 5 个链接，单击“开始”按钮即可下载，如图 2.36 所示。注意下载的时间与网络速度和文件量相关，有时候可能需要较长的时间下载。

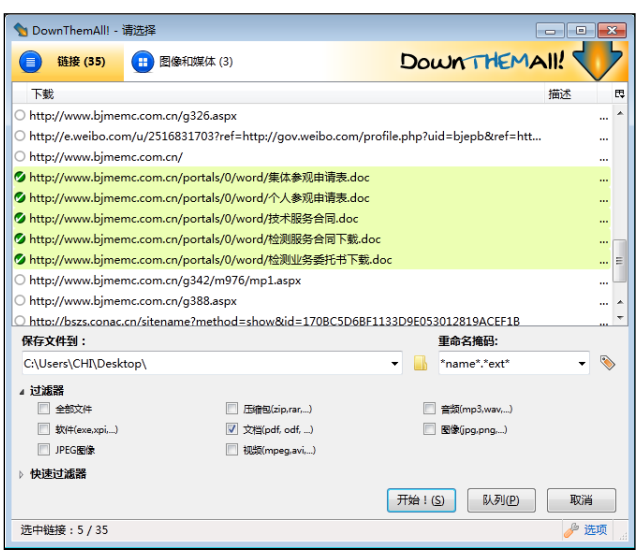


图 2.36　使用“DownThemAll”插件下载文件

国产软件迅雷也有类似的功能。首先登录迅雷[1]官方网站，下载并安装。在浏览器中（本案例使用的是 360 安全浏览器）登录北京市环境保护监测中心网站后用鼠标右键单击该页面，然后在打开的快捷菜单中选择【使用迅雷下载全部链接】，在打开的对话框的“任务类型过滤”中勾选“doc”复选框，确定下载文件保存地址，单击“立即下载”按钮即可，如图 2.37 所示。

图 2.37　使用迅雷下载文件

2.6 综合案例

2.6.1　使用联合国数据库

本案例介绍使用联合国数据库搜索信息、筛选信息、修改数据查看方式并将信息下载的方法。

（1）搜索数据。登录联合国数据库网址 http://data.un.org/，在搜索页面搜索“total fertility rate”，搜索结果如图 2.38 所示。单击搜索到的第一个结果，进入数据库界面。主界面显示“36534 records | Page 1 of 731”，即本次搜到的数据共 36 534 条，每页显示 50 行，共 731 页。

（2）筛选数据。使用页面左侧的筛选器筛选数据，在筛选器“Year（s）（30）”中分别勾选“1955-1960”……“2005-2010”共 11 个筛选项，如图 2.39 所示。单击“Apply Filters”，主界面显示“2820 records | Page 1 of 57”，即筛选后的数据共 2820 条。

1　http://www.xunlei.com/。

图 2.38　搜索联合国数据库

（3）修改数据查看方式。查看数据可以发现，数据行数太多查看不方便，可以将筛选项“Year（s）”显示在列。单击数据表上方的“Select the pivot column”，设置“Year（s）”为列头，如图 2.40 所示，单击“Update”更新数据显示方式。主界面显示“235 records | Page 1 of 5”，即筛选后的数据共 235 条。如图 2.41 所示显示了主界面的前 8 条数据。

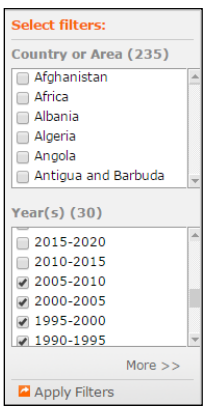


图 2.39　筛选器

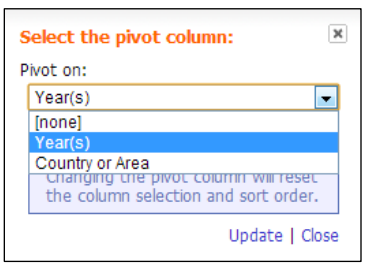


图 2.40　修改数据查看方式

Country or Area	Variant	1950-1955	1955-1960	1960-1965	1965-1970	1970-1975	1975-1980	1980-1985	1985-1990	1990-1995	1995-2000	2000-2005	2005-2010
Afghanistan	Medium variant	8	8	8	8	8	8	8	8	8	8	7	6
Africa	Medium variant	7	7	7	7	7	7	6	6	6	5	5	5
Albania	Medium variant	6	6	6	5	5	4	3	3	3	3	2	2
Algeria	Medium variant	8	8	8	8	8	7	6	5	4	3	2	3
Angola	Medium variant	7	7	7	7	7	7	7	7	7	7	7	7
Antigua and Barbuda	Medium variant	5	5	4	4	3	2	2	2	2	2	2	2
Areas not elsewhere specified	Medium variant	7	7	8	8	8	7	7	6	6	6	5	4
Argentina	Medium variant	3	3	3	3	3	3	3	3	3	3	2	2

图 2.41　显示前 8 条数据

因为所有记录的“Variant”值均相同，所以可以重新修改数据查看方式，增加或删除列。单击数据表上方的“Select columns to be displayed”，设置显示的列，勾选“Country or Area Code”，即增加该列，取消勾选“Variant”，即删除该列，如图 2.42 所示。单击“Update”更新数据显示方式。

单击数据表上方的“Select sort order”，设置排序方式按“Country or Area Code”升序排列，如图 2.43 所示。单击“Update”更新数据显示方式。

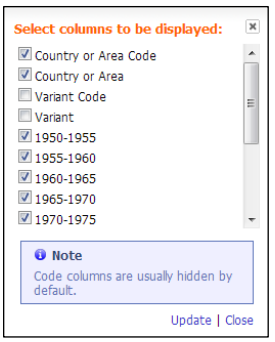


图 2.42　选择显示的列

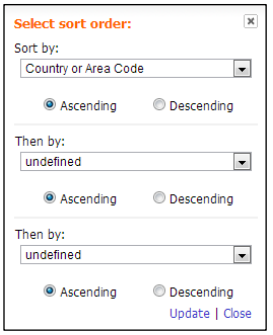


图 2.43　排序

设置完成后，前 8 行数据的前 5 列如图 2.44 所示。

Country or Area Code	Country or Area	1950-1955	1955-1960	1960-1965
4	Afghanistan	8	8	8
8	Albania	6	6	6
12	Algeria	8	8	8
24	Angola	7	7	7
28	Antigua and Barbuda	5	5	4
31	Azerbaijan	5	5	6
32	Argentina	3	3	3
36	Australia	3	3	3

图 2.44　前 8 行数据的前 5 列

（4）导出数据。单击数据表上方的“select download format”，选择导出格式是“Comma”，如图 2.45 所示。本列中导出的数据默认名称是“UNdata_Export_20151229_025626315.zip”（注意文件默认名称与导出的日期相关），用 Excel 打开解压的 CSV 文件，查看数据。

注意，图 2.44 中显示“Afghanistan”在“1950-1955”年的出生率是“8”，但使用 Excel 打开解压的 CSV 文件，这个数字是 7.6706，如图 2.46 所示。可以根据需要适当地减少小数位数。

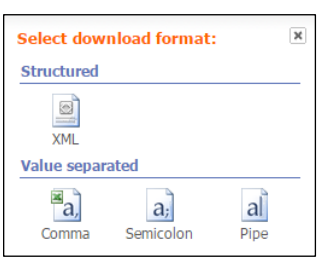


图 2.45　设置导出格式

B	C
Country or	1950-1955
Afghanistar	7.6706
Albania	6.112
Algeria	7.648
Angola	7
Antigua and	4.5
Azerbaijan	5.493
Argentina	3.154
Australia	3.18

图 2.46　Excel 打开效果

2.6.2　获取北京市 2014 年常住人口数量

获取北京市 2014 年常住人口数量最常见的方法是使用搜索引擎查询，如在百度页面中搜索“北

京市常住人口”，搜索结果如图 2.47 所示。

图 2.47　百度搜索结果

在搜索结果中可以查看到“2014 年末北京常住人口 2151.6 万人”，但这个数据是不能直接使用的，要找到数据的出处以确认数据的真实性。进入百度百科查看到此数据无出处，但可以查看到 2015 年数据来源于中国经济网 [1]，如图 2.48 所示。

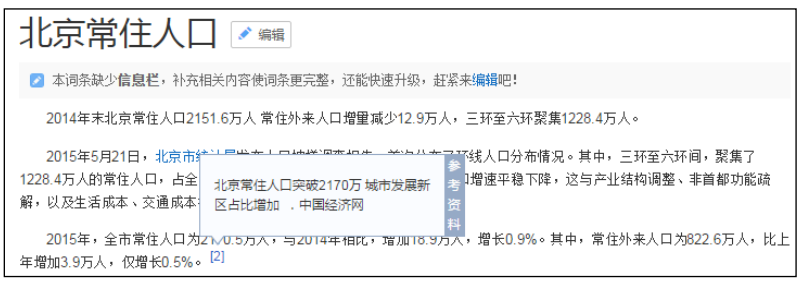
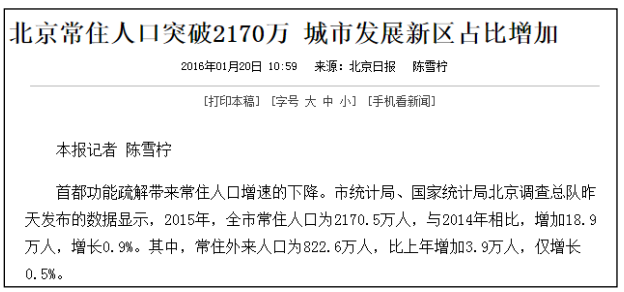


图 2.48　百度百科“北京常住人口”词条

单击数据来源链接到中国经济网，查看数据原始 Web 页面，如图 2.49 所示。查看到数据的出处是“市统计局、国家统计局北京调查总队”。


北京常住人口突破2170万 城市发展新区占比增加

2016年01月20日 10:59　来源：北京日报　陈雪柠

[打印本稿] [字号 大 中 小] [手机看新闻]

本报记者 陈雪柠

首都功能疏解带来常住人口增速的下降。市统计局、国家统计局北京调查总队昨天发布的数据显示，2015年，全市常住人口为2170.5万人，与2014年相比，增加18.9万人，增长0.9%。其中，常住外来人口为822.6万人，比上年增加3.9万人，仅增长0.5%。

图 2.49　中国经济网相关页面

1　数据来源 http://www.ce.cn/xwzx/gnsz/gdxw/201601/20/t20160120_8396634.shtml。

打开百度搜索引擎，搜索“北京市统计局”，获取其官方网址 www.bjstats.gov.cn。

打开北京市统计局网站，进入导航栏“统计数据”，因为 2014 年北京市常住人口是历史统计数据，2014 年的数据要 2015 年才能统计完成，所以在“年度数据”中选择“2015 年”，单击“查询”按钮。

在左侧的“人口与就业”导航栏中选择“常住人口总量（按区县分）（2014 年）”，显示常住人口 2151.6 万人，如图 2.50 所示。

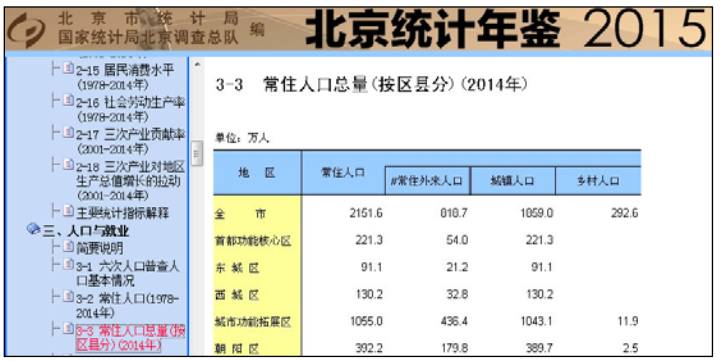


图 2.50　北京市统计局网站相关页面

此数据“2014 年北京市常住人口 2151.6 万人”可以在数据新闻中使用，使用时要脚注说明数据来源是“http://www.bjstats.gov.cn/nj/main/2015-tjnj/zk/indexch.htm”。

第 3 章

清理和分析数据

- “脏数据”（Dirty Data）
- 数据清理/分析工具
- 清理“脏数据”
- 使用 Excel 简单分析数据
- 数据清理原则
- 综合案例

3.1 “脏数据”（Dirty Data）

“多亏美国有着总体来说完善的公共档案法规，得到数据并不像在某些国家一样是个很大的问题。但是我们得到这些数据后，仍然面临着处理数据的一些麻烦，这些麻烦并非分析技巧上的，而是官僚系统带来的。这些数据往往是‘脏’的，大部分都是不标准的。有几次我得到的数据和它应该有的数据格式并不相符，也没有数据字典可供参考。有些机构仍然坚持发布尴尬的类似 PDF 格式的数据，还需要重新转换格式。这些问题让你在偶尔得到一些干净整洁的数据时会十分欣喜。”

——史蒂夫·多伊格（Steve Doig），沃尔特·克朗凯特新闻和大众传播学院，亚利桑那州立大学

在调查采访时，我们希望适当地运用提问技巧和方法，得到访问对象提供的正确信息。在数据新闻制作过程中，我们要诠释数据，通过数据分析得到结论。数据正确与否对结论起着决定性的作用。

在制作数据新闻时，一个非常重要但常常被遗忘的步骤就是清理原始数据。当我们通过多种方式和渠道获取数据时，这些数据往往是不能直接分析和使用的，需要进行预处理，即清理数据。因为原始数据存在各种各样的问题，如篡改数据、数据不完整、数据不一致、数据重复、数据存在错误、异常数据等，这些情况我们通称为存在“脏数据”。“脏数据”的存在不仅浪费时间，而且可能导致最终分析有误。

3.1.1 “脏数据”的成因

篡改数据。这是“脏数据”中最糟糕的一种形式，因为篡改数据是非常难以发现的。有些时候为了一些特定的利益，人们甚至会主动弄“脏”数据。例如，淘宝网卖家的相关数据，“刷信用”和“刷钻”等，导致淘宝网后台统计的是卖家篡改后的数据。

数据不完整。有些时候，数据的获取是比较困难的，如获取某个山区的 PM2.5 值，可能由于没有相关设备而导致该山区的数据缺失。我们期望数据符合完整性（Data Integrity）要求，即数据符合精确性（Accuracy）和可靠性（Reliability）。数据完整性包括实体完整性（Entity Integrity）、域完整性（Domain Integrity）、参照完整性（Referential Integrity）和用户自定义完整性（User-defined Integrity）。实体完整性是指一个关系中所有主属性（主码属性或标识性属性）不能取空值，即不能存在“空值”。域完整性是保证表中不能输入无效的值，如年龄是 90.5，这不符合常识（一般年龄是整数）。参照完整性一般应用于两个或两个以上的表，当在一个表中更新、删除或插入数据时，通过参照引用相互关联的另一个表中的数据，来检查对表的数据操作是否正确。用户自定义完整性是针对用户自定义的约束条件，如学生成绩范围是[0,100]，若存在某位学生的成绩是 120 的情况，则不符合用户自定义完整性。

数据不一致。数据的获取可能来自于不同的渠道，从而出现两份数据不一致的问题。如 2013 年 12 月 2 日，广州市政协委员韩志鹏分别向广东省卫计委、省财政厅和省审计厅快递了《关于公开广

东省 2012 年度社会抚养费收支及审计情况的申请》。12 月 4 日，广东省卫计委公布，2012 年度广东省社会抚养费征收总金额是 14.56 亿元。12 月 25 日，广东省财政厅答复，2012 年广东各地征收社会抚养费总额是 26.13 亿元。不同部门的审计结果相差了 11.57 亿元 [1]。数据的不一致也可能是由于重复存放的数据未能进行一致性地更新造成的。如各地最低工资标准在不同年代是不同的，如果来自不同渠道的数据存在重复且不一致的情况，可能是数据来源时间不同，也有可能是某个部门的数据调整而另一个部门没有及时更新导致的，如教师工资的调整，可能人事处的数据已经更新，而财务处的数据没有更新，就会产生矛盾的数据。

数据重复。由于数据来源的问题，可能存在一个数据记录两次的情况。要删除冗余的备份数据，确保同样的数据信息只被保存一次。

数据存在错误。这是“脏数据”中最糟糕的一种形式，主要是人为原因错误记录了信息，如工资是 6500.00 元，误记为 5600.00 元。2006 年美国国会选举期间，政府工作志愿者通过电话让已登记的选民来投票的过程中发现，有已经去世的选民依旧存在于登记表中，这也是数据存在错误的一种表现。

异常数据。异常数据是指某个数据与其他数据相比特别大或特别小，如获取的数据中教师的月薪大部分是几千元左右，若有某位教师的工资是 100 万元，则可以认为这是异常数据，可能这个异常数据是正确的，但对分析整体数据而言意义不大。

3.1.2 “脏数据”的表现形式

拼写问题。数据来源复杂，我们获得的数据可能存在数据格式不对的情况。如“性别”字段，可能包含各种各样的数值信息，如“男性”、“女性”、“男”、“女”、“1”、“0”、“男人”、“女人”、“F”、“M”、“Female”和“Male”等，甚至可能存在“男姓”或“Femal”这样的错误拼写。再如，职业字段中可以使用“Lawyer”、“Attorney”、“Atty”、“Counsel”或“Trial Lawyer”表示律师。类似的同一个名字可能有不同的拼写方式，如我国经常使用“中国”、“中华人民共和国”、“China”、“china”、“the People's Republic of China”或“PRC”等表示。这类问题的解决方法是统一成一种数值表示，如统一用“男”、“女”、“Lawyer”和“中国”表示。如果数据预处理时没有发现这类问题，在统计时带来的后果是计数错误。

数据格式问题。如“200”、“200.00”、“$200”、“20billion”、“20million”、“USD200”和“￥200”均表示金额，但是数据分析时有些数据是不合格的，如“20billion”、“20million”和“USD200”可能被计算机理解为字符而非数字。“$200”和“￥200”单位不同，一个是美金，另一个是人民币，应统一格式。“200”和“200.00”的精确度不一致，前者是整数，后者有两位小数，应统一精度。

1　人民网 http://finance.people.com.cn/n/2013/1226/c1004-23949709.html。

3.2 数据清理/分析工具

“脏数据”的成因和表现形式多种多样，我们必须要先清理“脏数据”，再做数据分析。数据清理（data cleaning），也称数据清洗，是指发现并纠正数据中的错误，要按照一定的规则“洗掉”存在的“脏数据”，包括检查数据一致性、处理无效值和缺失值、修改拼写问题和统一数据格式等。数据清理是一个反复的过程，不可能在短时间内完成，在一个数据新闻项目里，通常一半以上的时间都会被用在清理数据上，好数据是好新闻的基础。

清理“脏数据”的方式主要分为两种，一是手动清理，二是借助工具清理。前者适用于数据量较小的情况，后者适用于数据量较大的情况。合理选择数据清理和分析工具可以快速地修改或删除“脏数据”，多个工具的共同使用能发挥每个工具的优势，更好地节省时间。常见的数据清理及分析工具如下。

1．Excel

Excel 功能较强，简单易学。内嵌的各种函数帮助用户快速清除并修改数据，而且可以使用筛选、排序、分类汇总、数据透视表和图表等工具快速查看数据规律。Excel 还支持 VBA 编程，可以编写代码实现复杂的数据运算和清理工作。

2．OpenRefine

OpenRefine（以前的名称为 Google Refine）是一个免费、开源的数据清理工具，是谷歌公司专门为数据新闻工作者而开发的。它专注于清理杂乱数据，即使数据不太结构化也可以轻松快速探索大型数据集。OpenRefine 可以清理掉不必要的 HTML 代码、移除多余的符号和空格、连接多组数据和大小写转换等，并且可以输出为多种格式（如 CSV 和 XLS 等）。OpenRefine 界面与 Excel 类似，但工作方式更接近于数据库，这意味着它的功能比 Excel 更强大。

作为一个强大的数据处理和分析工具，OpenRefine 易学易用，而且数据可以保留在本地计算机上，这是敏感数据的福音。

3．R 语言

R是用于数据处理和绘图的语言。它是属于GNU[1]计划的一个自由、免费而且源代码开放的软件。它优秀的统计制图功能，以及可操纵数据的输入和输出、分支、循环和用户可自定义功能使其逐渐被应用到数据新闻中，比如《纽约时报》的“512 条通往白宫的路”，就是利用统计里的决策树模型[2]。

1 GNU 计划，又称革奴计划，是由 Richard Stallman 在 1983 年 9 月 27 日公开发起的。它的目标是创建一套完全自由的操作系统。

2 http://www.nytimes.com/interactive/2012/11/02/us/politics/paths-to-the-white-house.html?_r=0。

4．Data Wrangler

Data Wrangler（网址 http://vis.stanford.edu/wrangler/）是由斯坦福大学开发的一个在线的数据清理和转换工具，可以减少用户格式化数据的时间。Data Wrangler 免费而且简单易学，但它是基于网络的服务，数据必须上传到外部网站，不适合清理敏感数据。

5．Python

Python 是数据获取、数据清理和数据挖掘时经常使用的语言。Python 是免费的，而且以语法简捷著称，代码易读而且可扩展性强，数据挖掘包多且安装方便，常见库包括 Python 标准库、Numpy 与 Scipy、Matplotlib 和 Scikit Learn 等。如果使用数据收集工具无法获得需要的数据，则许多记者使用 Python 编写爬虫程序获取数据。

3.3 清理“脏数据”

本节以 Windows 操作系统下的 OpenRefine 为例，介绍“脏数据”清理工具的安装方法、数据导入和导出、数据归类和数据筛选、单元格和行列的编辑、变换和排序等功能。为了更好地使用 OpenRefine，本节还介绍了常见函数的使用说明，最后详细介绍了正则表达式的使用。本节大量的案例意在呈现数据清理的思路和方法。

3.3.1　安装 OpenRefine 环境

可以到 OpenRefine 官网 http://openrefine.org 下载最新版本，具体下载地址是 http://openrefine.org/download.html，最新（2016 年 5 月）的版本是 OpenRefine 2.6。OpenRefine 相关的文档信息资源可到 http://openrefine.org/documentation.html 下载，目前只有英文文档。

OpenRefine 基于 Java 环境开发，因此是跨平台的，可以安装在 Linux、Windows 和 MAC 等操作系统。安装 OpenRefine 时若操作系统没有安装 Java，将自动转到 Java 官网安装所需程序。也可以在安装 OpenRefine 前手动安装 Java 环境，安装包在 Java 官网下载（http://www.java.com/zh_CN/），可以根据操作系统下载相应的文件并安装，如图 3.1 所示。

以 Windows 操作系统为例，不仅分为联机版本和脱机版本两种，还有 32 位浏览器和 64 位浏览器之分，需要下载相应的 Java 版本安装。建议删除操作系统中旧的 Java 版本并安装最新的版本。从 Java 8 Update 20（8u20）开始，在 Windows 操作系统上，Java 卸载工具已经集成在安装程序中，用于从操作系统中删除较早的 Java 版本。

图 3.1　Java 下载安装包

在 Windows 系统中安装 OpenRefine 时，首先下载 RAR 压缩包，OpenRefine 2.6 下载后是一个仅有 38MB 的压缩包，解压后，双击解压文件夹中的 openrefine.exe 文件即可，如图 3.2 所示。

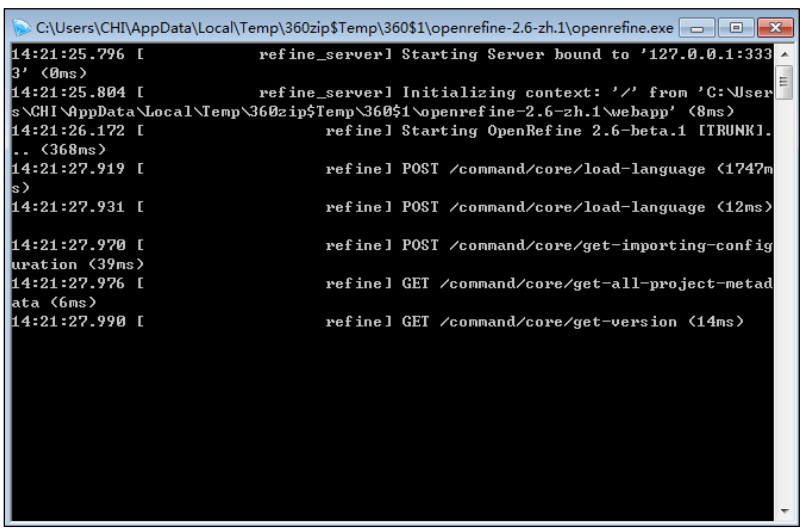


图 3.2　Windows 系统下安装 OpenRefine

在 Mac 系统中安装 OpenRefine 时，首先下载 DMG 压缩包，打开磁盘镜像，拖动 OpenRefine 的图标到 Applications 目录即可。

在 Linux 系统中安装 OpenRefine 时，首先下载 GZ 压缩包，解压到当前用户的 home 目录，在终端命令行环境输入“./refine”以启动 OpenRefine。

3.3.2　创建项目（导入数据）

OpenRefine 支持的文件格式包含以下几种。

- 逗号分隔值（CSV）、制表符分隔值（TSV）和其他“*SV”格式。
- Excel 文档（包括 XLS 和 XLSX）和开放文档格式（ODF）的电子表格（.ods）。
- JavaScript 对象符号（JSON）、XML 和 XML 资源描述框架（RDF）。
- 基于行的格式（日志）。

如果导入其他格式的文件，可以通过 OpenRefine 扩展的方式导入。

要运行 OpenRefine，可双击解压文件夹中的 openrefine.exe 文件，默认浏览器在 URL“http://127.0.0.1:3333/”或者“http://localhost:3333/”打开 OpenRefine。

OpenRefine 运行界面的左侧包含四个选项卡，分别是“新建项目”、“打开项目”、“导入项目”和“语言设定”，如图 3.3 所示。

图 3.3　OpenRefine 运行界面

首次使用 OpenRefine，一般选择“新建项目”。通过“新建项目”设置数据来源，即将数据加载到 OpenRefine。数据来源包括四种，使用“这台电脑”选项可以选择存储在本地计算机的数据文件，使用“网址（URLs）”选项可以导入一个或多个下载数据的来源网址。“剪贴板”选项用于将复制和粘贴的数据转换为文本字段。使用“Google Data（谷歌数据）”选项允许访问谷歌电子表格或 Fusion 表（需要互联网连接）。

如果非首次使用 OpenRefine，可以选择“新建项目”或“打开项目”。“打开项目”用于打开以前建立的项目。OpenRefine 按时间倒序的方式显示现有项目列表。

如果存储了 OpenRefine 的项目文件（.tar or .tar.gz），可以选择“导入项目”直接导入这个项目。

“配置解析选项”包含两个部分，如图 3.4 所示，在左侧选择“数据解析格式”后，右侧是格式对应的设置选项，如将“字符编码”设置为“UTF-8”以正确显示中文数据（否则可能是乱码）。设置“数据中列的分隔方式”是逗号、制表符或自定义符号，“忽略文件首部的前__行”表示忽略数据源的前几行数据。“将其次的下_line（s）作为列头”表示下几行是列标题。

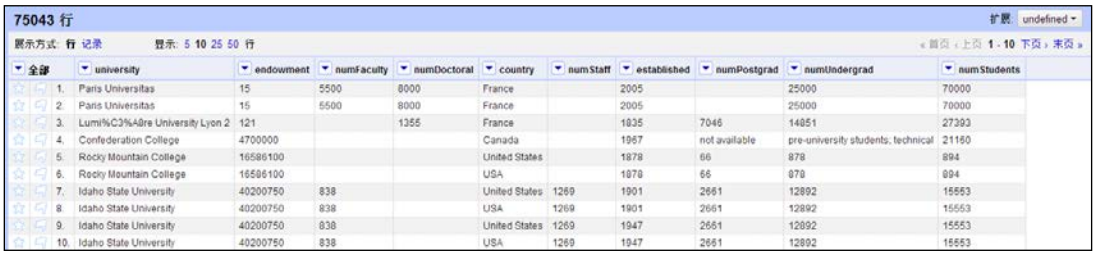
图 3.4　OpenRefine 配置解析选项

在右上角输入“项目名称”后单击“新建项目”按钮即可新建一个项目，如图 3.5 所示。

图 3.5　OpenRefine 新建项目

3.3.3　主界面

数据加载后进入 OpenRefine 主界面，可以按照多种不同的方式查看数据，如图 3.6 所示。



全部	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
1.	Paris Universitas	15	5500	8000	France		2005		25000	70000
2.	Paris Universitas	15	5500	8000	France		2005		25000	70000
3.	Lumi%C3%A8re University Lyon 2	121		1355	France		1835	7046	14851	27393
4.	Confederation College	4700000			Canada		1967	not available	pre-university students, technical	21160
5.	Rocky Mountain College	16586100			United States		1878	66	878	894
6.	Rocky Mountain College	16586100			USA		1878	66	878	894
7.	Idaho State University	40200750	838		United States	1269	1901	2661	12892	15553
8.	Idaho State University	40200750	838		USA	1269	1901	2661	12892	15553
9.	Idaho State University	40200750	838		United States	1269	1947	2661	12892	15553
10.	Idaho State University	40200750	838		USA	1269	1947	2661	12892	15553

图 3.6　OpenRefine 主界面

图 3.6 的上方显示数据的总行数(共包含 75043 行数据)、各种显示选项(每页显示 10 行数据)、选择页数、列标题和菜单及实际的单元格内容。

OpenRefine 的显示选项共 4 种，分别是每页显示 5、10、25 或 50 行记录。

主界面中每列均有一个菜单，可以通过单击列标题左侧的三角形下拉按钮进行选择，如图 3.7 所示。使用菜单可以对该列进行归类、筛选、编辑、变换和排序等操作。

数据清理前通过 OpenRefine 主界面认真查看数据是非常重要的，如查看每列的数据类型和格式是否正确、单元格是否有空值等。

图 3.7　“归类”菜单

3.3.4　归类（Facet）

归类是 OpenRefine 最常使用的功能之一，归类并不影响数据的值，只是一种查看数据的方式，归类有很多种，参见图 3.7。其中“文本归类”、“数值归类”和“自定义归类”使用的较为频繁。

1．文本归类（Text Facets）

“文本归类”是将所选的列按照文本规则分类汇总，类似于 Excel 的筛选和分类汇总。“文本归类”适用于文本种类不是太多的情况，如果文本的种类有几百上千种，查看是不可控的，效果不佳。

如数据中包含“country”列，可以通过“文本归类”查看每个国家的记录数。单击“country”列左边的三角形下拉按钮，选择【归类】|【文本归类】(【Facet 】|【Text Facet】) 选项，“country”列的分组结果如图 3.8 和图 3.9 所示。

这种方式特别适合查看数据中元素的分布情况，本例中包含 68 个“country”值，图 3.8 按“名称”升序的方式显示每个国家的记录条数，如“Albania”共 8 条记录。图 3.9 按“数量”降序的方式显示每个国家的记录条数，如“USA”共 6401 条记录。

图 3.8　按“名称”升序显示

图 3.9　按“数量”降序显示

Cluster（聚类）用于对相似的值进行聚类分析，方便用户查找“脏数据”。单击“聚类”按钮，在“簇集&编辑列‘country’”窗口查看到共有 3 个簇集，如第一个簇集的大小是 2，行数是 6794，簇中值分别是“USA（6401 rows）”和“U.S.A.（393 rows）”，右侧是图形化的“簇中的行数”，如图 3.10 所示。

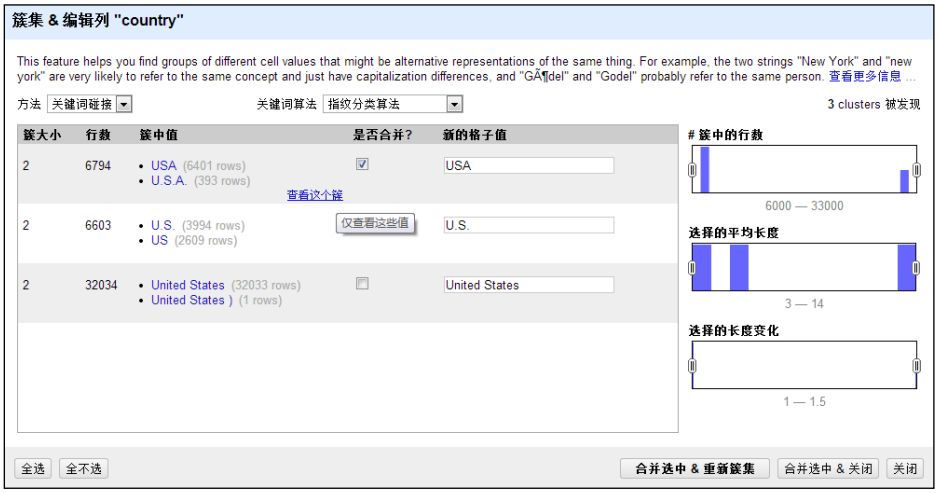
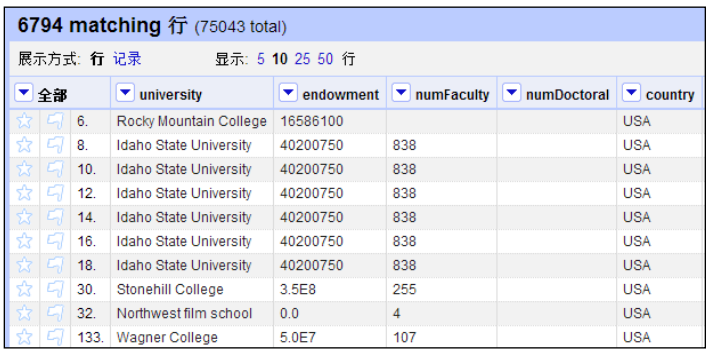


图 3.10　显示“country”簇集

单击“查看这个簇”链接，将在新的浏览器窗口中显示这个簇集的 6794 条记录，如图 3.11 所示。注意“country”列的值一定是“USA”或“U.S.A.”，左上角显示数据集共有 75043 行，当前匹配记录共 6794 条。

图 3.11 簇集中的“USA”和“U.S.A.”均是美国的表示方法，同一个信息用多种方法表示是典型的“脏数据”，勾选“是否合并”复选框，并确认“新的格子值”是“USA”以合并簇集。这时在“簇集&编辑列‘country’”窗口中查看到只有两个簇集，如图 3.12 所示。合并簇集后归类由 68 个“country”值变为 67 个。



全部	university	endowment	numFaculty	numDoctoral	country
6.	Rocky Mountain College	16586100			USA
8.	Idaho State University	40200750	838		USA
10.	Idaho State University	40200750	838		USA
12.	Idaho State University	40200750	838		USA
14.	Idaho State University	40200750	838		USA
16.	Idaho State University	40200750	838		USA
18.	Idaho State University	40200750	838		USA
30.	Stonehill College	3.5E8	255		USA
32.	Northwest film school	0.0	4		USA
133.	Wagner College	5.0E7	107		USA

图 3.11　查看某个簇集

同样可以合并其他两个簇集（注意，确认“新的格子值”的内容均是“USA”），合并后显示包含 63 个“country”值。如上操作后，“USA”、“U.S.A.”、“U.S.”、“US” 、“United States”和“(United States)”均合并为“USA”，分类减少了 5 个。

簇集有两种方法生成，关键词算法也包含多种。关键词算法不同，生成的簇集也不同。如本例中使用的方法是“关键词碰接”，使用的关键词算法是“指纹分类算法”，如图 3.12 所示。

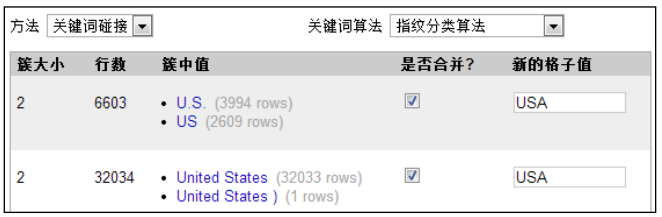


图 3.12　合并簇集

2．多重归类（Multiple Facets）

可以创建多个归类，多个归类同时使用可以更精准地查看数据。此功能类似于 Excel 的多条件筛选。

如对“established”列新建“文本归类”，显示包含 308 个“established”值，如图 3.13 所示。在“country”文本归类中选择“USA”，在“established”文本归类中选择“1933”，则主界面显示仅有 34 条记录匹配，如图 3.14 所示。

使用同样的方法可以创建更多的“文本归类”、“数值归类”和“自定义归类”等实现多重归类。

图 3.13　设置多重归类

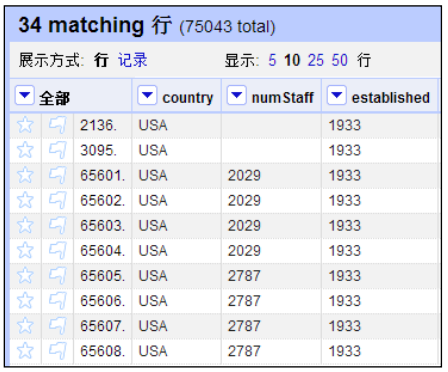


图 3.14　多重归类匹配记录

3．数值归类/日期归类（Numeric Facets/Date Facets）

在 OpenRefine 中，数值型数据右对齐并呈绿色显示（文本默认左对齐并呈黑色显示），如图 3.15 所示。单击“PM2.5”列左边的三角形下拉按钮，选择【归类】|【数值归类】（【Facet 】|【Numeric Facet 】)选项，将数值按范围归类。数值归类用于对数值型数据进行归类，可以更客观地查看列信息，

如图 3.16 所示。

全部	城市	PM2.5
1.	海口	34
2.	南昌	54
3.	长沙	69
4.	南宁	53
5.	衢州	47
6.	贵阳	43
7.	佛山	53
8.	成都	63
9.	武汉	71
10.	珠海	41

图 3.15　数值型数据 PM2.5

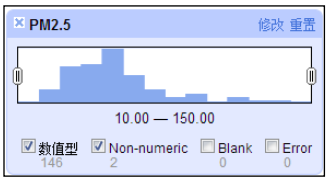


图 3.16　数值归类 PM2.5

图 3.16 中显示，记录共 148 条，其中数值型数据 146 条，非数值型数据 2 条，无空记录和错误记录。PM2.5 的数值范围是 10.00 到 150.00。

仅勾选“Non-numeric”复选框，如图 3.17 所示。在主界面查看 2 条非数值型记录，如图 3.18 所示。这两条记录的 PM2.5 值均是左对齐黑色显示，即为文本而非数值型数据，因为数据中包含符号“+”和文字“大约”。

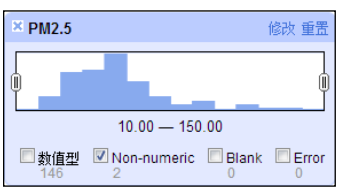


图 3.17　勾选“非数值”

图 3.18　显示非数值型记录

单击“edit”编辑单元格，将“149+”数据类型修改为“数字”，值为“149”，单击“应用”按钮，如图 3.19 所示。用同样的方法修改另一条记录，编辑后的记录显示效果如图 3.20 所示。

图 3.19　编辑单元格

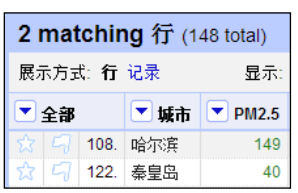


图 3.20　编辑后的记录

单击“刷新”按钮，查看刷新后的数值归类，如图 3.21 所示。

日期归类与数值归类类似，日期型数据也是右对齐并呈绿色显示。单击“日期”列左边的三角形下拉按钮，选择【归类】|【日期线归类】(【Facet 】|【Date Facet】) 选项，将日期按范围归类，如图 3.22 所示。

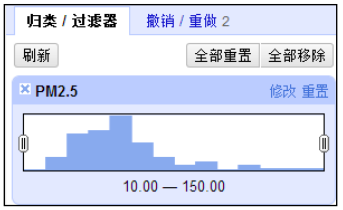


图 3.21　刷新后的归类

图 3.22　日期线归类

图 3.22 显示，记录共 148 条，其中时间数据 147 条，非时间数据 1 条，无空白记录和错误记录。日期是 2015-10-01。用编辑非数值型数据的方法编辑“非时间”记录，修改后的记录如图 3.23 所示。注意日期修改后的颜色变化。

图 3.23　修改后的记录

4．自定义归类（Customized facets）

自定义归类可以根据用户需要设置归类方式，包含“自定义文本归类”、“自定义数值归类”和“自定义归类”三个子菜单，前面两种方式需要输入表达式实现归类（具体内容参见 3.3.12 小节“函数”和 3.3.13 小节“正则表达式”）。

如使用“自定义文本归类”对城市名称中包含“海”的城市归类，则单击“城市”列左边的三角形下拉按钮，选择【归类】|【自定义文本归类】，在窗口中输入表达式 value.contains（“海”），如图 3.24 所示。注意，表达式中的符号均为英文。“预览”标签显示执行该表达式后的值，本例中城市名要么包含“海”，要么不包含，在“预览”中可以查看到，“海口”执行表达式后的结果是“true”，而“南昌”执行表达式后的结果是“false”。单击“确定”按钮对所有记录执行该表达式，在“城市归类”中查看执行表达式后的归为 2 类，“false”共 142 条记录，“true” 共 6 条记录，如图 3.25 所示。单击“true”，查看匹配的 6 条记录，如图 3.26 所示。

单击【全部】|【编辑行】|【移除所有匹配的行】选项将删除匹配的 6 条记录。单击【撤销/重做】中“Remove 6 rows”之前的操作，即可撤销删除 6 条记录的操作。

“自定义归类”包括“按字归类（Word facet）”、“复数归类（Duplicates facet）”、“数字对数归类（Numeric log facet）”、“约为 1 的数字对数归类（1-bounded numeric log facet）”、“文本长度归类（Text length facet）”、“文本长度的对数值归类（Log of text length facet）”、“Unicode 字符归类（Unicode char-code facet）”、“按错误归类（Facet by error）”和“按空白归类（Facet by blank）”共 9 种。

图 3.24 “自定义文本归类”的表达式

图 3.25 “自定义文本归类”

图 3.26 显示 6 个匹配的记录

如单击“城市”列左边的三角形下拉按钮，选择【归类】|【自定义归类】|【文本长度归类】选项，此时界面如图 3.27 所示，拖动滑块到“3.32”，则主界面仅显示城市名称的长度范围是“3.32 ~ 4.02”的记录，即城市名称是 4 个字的记录，如图 3.28 所示。

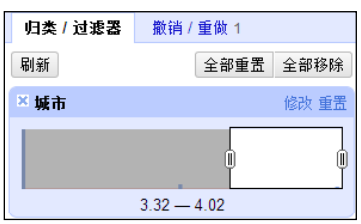


图 3.27 按“文本长度归类”

图 3.28 显示城市名称为 4 个字的记录

“按空白归类（Facet by blank）”也是一种经常使用的归类方式，该归类将选择列的记录分为“空值”和“非空值”两种。

为区分归类后的数据，经常使用“星标（Stars）”标明优质或感兴趣的记录，使用“标记（Flag）”标明坏记录，以方便后期查看。

如对城市名称中包含“海”的记录感兴趣，归类后，在“全部”列下拉菜单中选择【归类】|【加星标】，如图 3.29 所示。

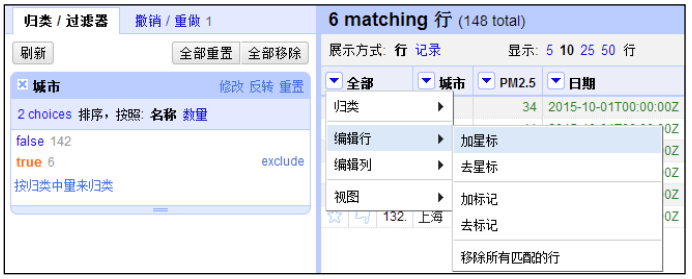


图 3.29　为归类记录加星标

“加星标”后的记录如图 3.30 所示。单击“城市归类”窗口中“true”后面的“exclude”，显示全部记录，如图 3.31 所示。可以查看“加星标”的效果，注意，本例中每页显示 10 行，首页“加星标”的记录仅有 2 条。

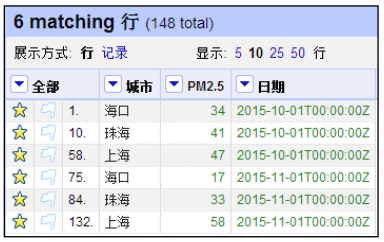


图 3.30　“加星标”后的记录

图 3.31　显示加星标后所有城市名称的记录

在“全部”列下拉菜单中选项【归类】|【去星标】，可以删除“星标”。添加和删除“标记”的方法与“星标”的操作相同。

“复数归类（Duplicates facet）”也是一种经常使用的操作，该归类判断记录值是否有重复，显示“重复”和“非重复”的记录条数。

3.3.5　文本过滤器（Text filter）

文本过滤器可以在特定的列中筛选包含某些精确字符串的单元格，或者匹配某些正则表达式的单元格。如在“country”列选择【文本过滤器】，输入要筛选的文本内容或者正则表达式，如输入“USA”并勾选“大小写敏感”复选框，如图 3.32 所示，则仅显示符合条件的数据。

使用正则表达式也可以对某列进行过滤，如搜索以字母“C”开头并以数字结尾的单元格，输入的正则表达式是“^[C].*\d$”，勾选“正则表达式”复选框，如图 3.33 所示。注意，有可能需要单击“刷新”按钮刷新约束。正则表达式的详细内容参见 3.3.13 小节。

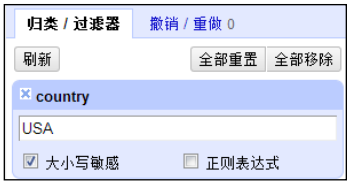


图 3.32 “字符串”文本过滤器

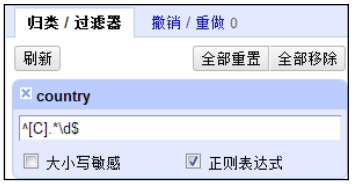


图 3.33 “正则表达式”文本过滤器

3.3.6 编辑单元格（Edit cells）

使用“编辑单元格”菜单可以实现单元格格式转换，如使用公式修改单元格中的数值，使用“常用转换”将文本转换为全部大写、将文本转换为数值等，还可以分离和合并多值单元格等，如图 3.34 所示。

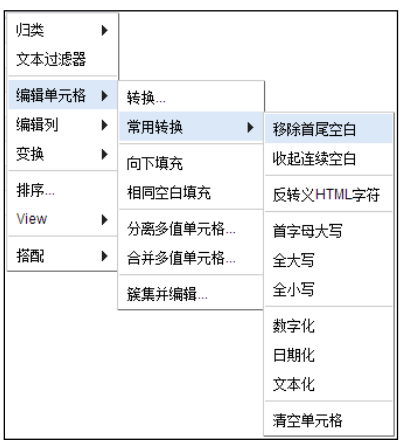


图 3.34 “编辑单元格”菜单

“常用转换”是使用频率很高的功能之一。如“移除首尾空白”类似于 Excel 的 Trim( )函数，即删除文本前面和后面的空格。也可以直接输入函数“value.trim( )”实现。

“收起连续空白”将字符串中的多个空格转换为 1 个空格，如“北 京”（中间包含 1 个空格）、“北 京”（中间包含 2 个空格）、“北 京”（中间包含 3 个空格），执行该操作后均转换为“北 京”（中间包含 1 个空格）。也可以直接输入函数“value.replace（/\s+/,' '）”实现。

“反转义 HTML 字符”适合处理包含超文本标记语言（HTML）的单元格，通过该操作转换特殊字符。

“数字化”、“日期化”和“文本化”是将原单元格格式转换为相应的格式。对应的三个函数分别是 value.toNumber( )、value.toDate( )和 value.toString( )。以“数字化”为例，如某列的原信息是“90.0”、“A+”、“87.0”、“90+”和“>90”，执行“数字化”操作后转换为“90”、“A+”、“87”、“90+”和“>90”，如图 3.35 所示。主界面显示“Text transform on 2 cells in column 成绩: value.toNumber( )”，即仅有 2

个单元格完成了转换。注意，并不是所有文本均可以转换为数值型数据，如本例中“A+”、“90+”和“>90”转换前后均是文本。

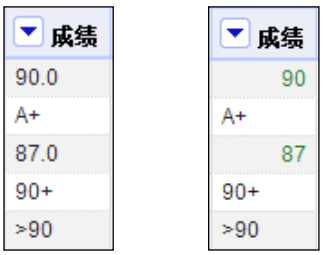


图 3.35　“数字化”操作前后的效果对比

“转换”功能也是非常有用的一项功能，但该功能要求用户熟练使用函数和表达式。如希望将“成绩”列的所有值转换为数值型数据，且分数变为原来的 100 倍，需要在“成绩”列下拉菜单中选择【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“value.toNumber( )*100”。注意，OpenRefine 区分大小写，与 Java 语言类似。预览中显示了表达式对各个记录的操作结果，可以转换为数值型数据的记录实现乘百操作，而其他记录则显示无法转换为数值型数据的错误信息，如图 3.36 所示。

图 3.36　自定义单元格内容转换

“分离多值单元格”功能帮助用户按照某种分隔符分割记录。如某个分类值是“手机和平板”或“桌子和椅子”，当用户想详细了解到底有多少种类时，可以用此功能分割记录值。

原始数据如图 3.37 所示，对“分类”列执行【编辑单元格】|【分离多值单元格】操作，在打开

的对话框中输入分隔符“和”（出现“和”的单元格被认为是多值单元格，执行分离操作），该列分离后的效果如图 3.38 所示。该功能方便用户按更详细的类别统计、归类等。

分类	数量	金额
电脑	1	6000
配件	12	1200
手机和平板	4	9400
桌子和椅子	6	600
电脑	2	8600
配件	8	420
桌子和椅子	2	2900
手机和平板	2	5200

图 3.37 原始数据

分类	数量	金额
电脑	1	6000
配件	12	1200
手机	4	9400
平板		
桌子	6	600
椅子		
电脑	2	8600
配件	8	420
桌子	2	2900
椅子		
手机	2	5200
平板		

图 3.38 “分离多值单元格”后的数据

“合并多值单元格”与“分离多值单元格”的功能正好相反。对“分类”列执行【编辑单元格】|【合并多值单元格】操作，在打开的对话框中输入分隔符“+”，如图 3.39 所示，则该列合并后的效果如图 3.40 所示。

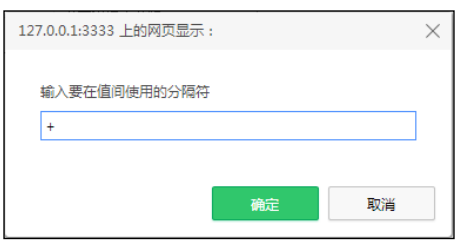


图 3.39 输入分隔符

分类	数量	金额
电脑	1	6000
配件	12	1200
手机+平板	4	9400
桌子+椅子	6	600
电脑	2	8600
配件	8	420
桌子+椅子	2	2900
手机+平板	2	5200

图 3.40 “合并多值单元格”后的数据

“清空单元格”用于清除单元格的内容，即单元格为 null。

“相同空白填充”是对相同内容的单元格填充空白，一般适用于删除重复记录，具体使用方法参见 3.6.2 小节的步骤 9。

3.3.7 编辑列（Edit column）

编辑列操作包含按分隔符或字段长度等方式分割列、重命名列、移除列，也可以将列移动到合适的位置，“编辑列”菜单如图 3.41 所示。

“分割此列”功能用于将列按照某种规则分割为多列，如按分隔符、正则表达式分割或按字段长度分割等，如图 3.42 所示。

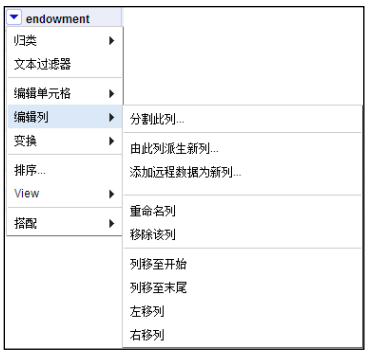


图 3.41　“编辑列”菜单

图 3.42　设置“分割此列”操作

“由此列派生新列”功能是根据已经存在的列生成新列，如根据 AQI（空气质量指数）判定空气质量，AQI 小于等于 50 则空气质量为“优”，AQI 在 50～100 之间则空气质量是“良”，AQI 在 100～200 之间则空气质量是“中度污染”，AQI 在 200 以上则空气质量是“重度污染”。首先输入新列名称“空气质量”，然后输入表达式“if（value<=50,"优",if（value<=100,"良",if（value<=200,"中度污染","重度污染"））)”，在“预览”选项卡中可以查看新建列的具体值，如图 3.43 所示。

“重命名列”功能可以重命名选择的列，“移除该列”功能将删除选择的列，单击“撤销”可以撤销这两种操作。

“列移至开始”、“列移至末尾”、“左移列”和“右移列”都是移动选择列的位置，方便用户查看数据。

图 3.43　设置“由此列派生新列”操作

3.3.8 变换（Transpose）

有时候数据行或数据列的显示方式并不是用户期望的样式，这时可以使用“变换”菜单修改数据的显示方式，如将行转换成列，或者将列转换成行，如图 3.44 所示。

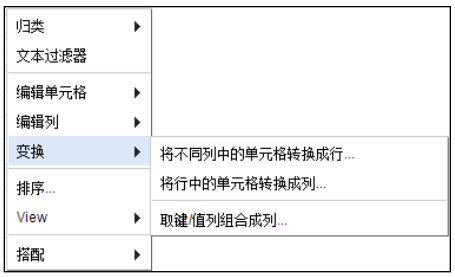


图 3.44 “变换”菜单

如图 3.45 所示的数据，每行记录中均有空值，因为为每种物品设置了 5 个度量属性，但有些度量值没有数据，如矩形的桌子没有“直径”值。

5 行

展示方式: 行 记录　　显示: 5 10 25 50 行

全部	名字	长	宽	高	直径	重里
1.	桌子	100	60	85		
2.	椅子	55	60			
3.	手机	12	6			260
4.	平板	7	8	9		320
5.	水杯				8	350

图 3.45 原始数据

选择某列，单击【变换】|【将不同列中的单元格转换成行】选项，可以将特定的某些列按照一个规则转换为行。在打开的对话框中，“来源列”用于设置开始转换的列，默认情况下在哪列操作即在哪列开始转换。“目的列”用于设置结束转换的列，即“来源列”和“目的列”中间的列执行转换操作。在“转换成”选项区中若选择“两个新列”，则输入的键列是主键（Key column），包含原始的列名，值列包含原始的记录值；若选择“一个列”，则键列和值列显示在一列，可以设置键列和值列的分隔符。勾选“忽略空白单元格”则仅显示无空白值的记录，勾选“向下填充其他单元格”则在键列中向下填充，否则只显示首个键列，其他键列值是空。

在“长”列中单击【变换】|【将不同列中的单元格转换成行】选项，在“来源列”中选择“长”，在“目的列”中选择“重量”，在“转换成”选项区中设置“一个列”的“属性”，勾选“将原始列的名称前缀给各个单元格添加‘:’在单元格值的前面”复选框，再勾选“忽略空白单元格”复选框。转换后生成 14 条记录，效果如图 3.46 所示。

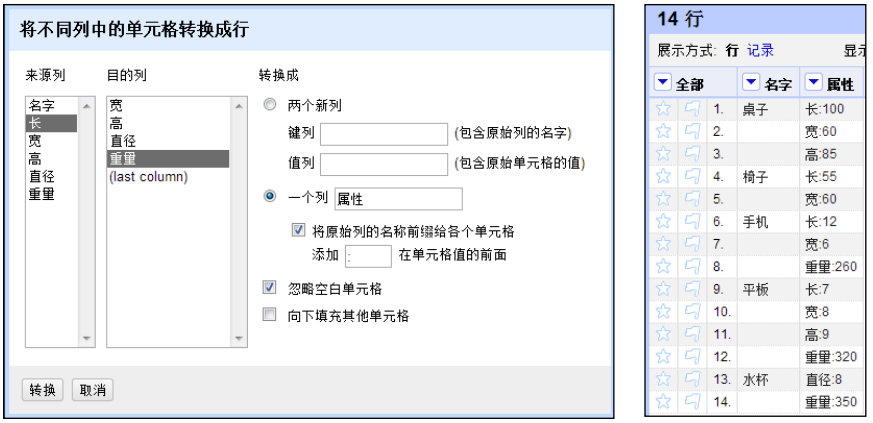


图 3.46　“转换成”选项区设置为“一个列”及其效果

撤销刚才的操作。在“长”列中单击【变换】|【将不同列中的单元格转换成行】选项，在“来源列”中选择“长”，在“目的列”中选择“重量”，在“转换成”选项区中设置“两个新列”，分别是键列“度量属性”和值列“度量值”，勾选“忽略空白单元格”和“向下填充其他单元格”复选框，转换后生成 14 条记录，如图 3.47 所示。对比图 3.46 和图 3.47 的异同。

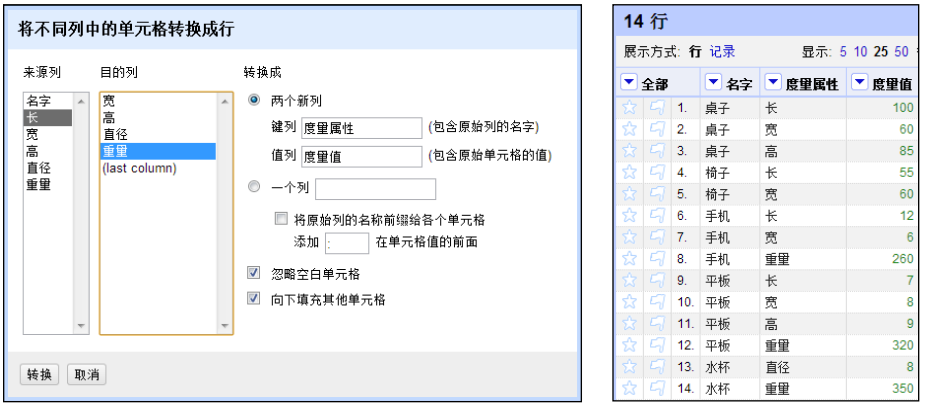


图 3.47　“转换成”选项区设置为“两个新列”及其效果

选择某列，单击【变换】|【取键/值列组合成列】选项完成“将不同列中的单元格转换成行”的相反操作，即将键列和值列转换为列，这种操作要小心使用，因为此操作对空白单元格敏感。如选择图 3.47 中的“度量属性”列，单击【变换】|【取键/值列组合成列】选项，在打开的对话框中设置键列为“度量属性”，设置值列为“度量值”，单击“确定”按钮后这两个列将组合成列，如图 3.48 所示。

注意记录与原始数据是有区别的（对比图 3.45 和图 3.48），图 3.48 中包含多个空白行，而且列的顺序可能发生了变化（注意“重量”和“直径”两列的顺序）。

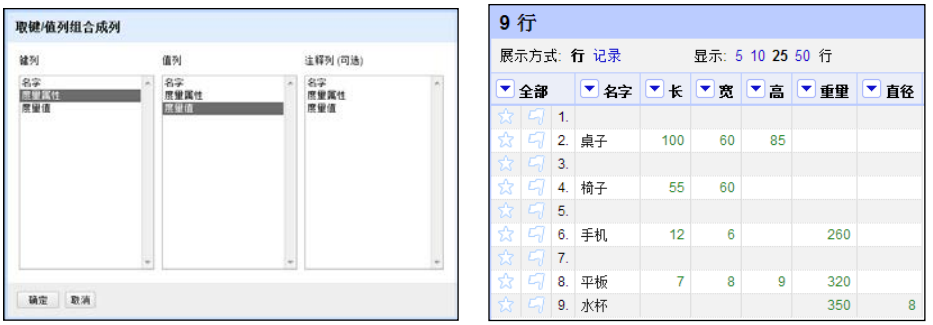


图 3.48 “取键/值列组合成列”操作及其效果

为回到原始数据的效果，按“名字”文本归类，选择“blank”后，单击【全部】|【编辑行】|【移除所有匹配的行】选项将删除匹配的 4 条记录。移除“名字”归类，显示 5 条记录。再选择“直径”列，单击【编辑列】|【左移列】选项即可。

3.3.9 排序（Sort）

选择列后，针对该列可以按照多种方式升序或降序排序。在排序对话框中，左侧可以设置排序的依据，右侧使用鼠标拖曳的方式确定“空白”、“错误”和“合法值”的排序方式。一般情况下将“空白”和“错误”排在“合法值”的前面，方便对这些非合法数据快速处理，如图 3.49 所示。

图 3.49 排序对话框

排序后的列可以重新排序、反转或不排序，也可以固定排序后行的顺序，如图 3.50 所示。重新排序类似于 Excel 自定义排序中设置多个条件的排序。

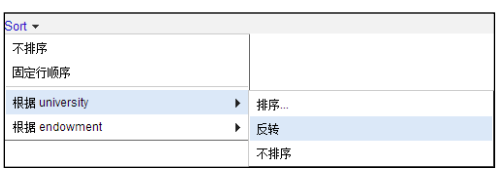


图 3.50 排序操作菜单

使用“撤销/重做”选项可以实现回退和前进操作，如图 3.51 所示。

图 3.51　“撤销/重做”选项

3.3.10　视图（View）

数据中的列可能非常多，为方便查看和操作，可以收起不关注的列，“view”菜单如图 3.52 所示。收起列的方式共有四种，分别是“收起该列”、“收起所有其他列”、“收起左侧列”和“收起右侧列”。

在列标题“全部”中可以展开所有列。单击【全部】|【视图】|【展开所有列】选项可以将收起的列全部展开，单击【全部】|【视图】|【收起所有列】选项将收起所有列，如图 3.53 所示。

图 3.52　“View”菜单

图 3.53　“全部”菜单

3.3.11　导出（Export）

使用主界面右上角的“导出”按钮可以导出项目，即将整个 OpenRefine 项目导出为一个.tar 或.gz 文件，这样就可以方便他人导入自己的 OpenRefine。导出的文件既包含数据，也包含数据的变化历史（包含撤销/重做）。导入项目文件的用户可以清楚地查阅操作历史，甚至可以撤销某一个或多个操作。“导出”菜单如图 3.54 所示。

图 3.54　“导出”菜单

“导出”按钮也可以仅导出数据。导出的数据格式包括以 tab 分隔的值、以逗号分隔的值、HTML 表格、Excel 和 ODF 电子表格。需要注意的是，如果使用了过滤器，则仅导出匹配的记录行和过滤器约束。

“Triple loader”和“MQLWrite”选项包含高级选项，用于通过 Freebase 扩展导出数据。

“自定义表格导出器”用于设置导出的具体内容，如仅导出选中的列、是否导出空白行、是否导出列头、是否排序列和设置时间格式等，如图 3.55 所示。

选择“正在生成模板”选项可以使用个性化模板输入所需格式，如将某列数据变更为 JSON（JavaScript Object Notation），这是一种轻量级的数据交换格式，行模板如下：

```
{
     "university" : {{jsonize(cells["university"].value)}},
     "endowment" : {{jsonize(cells["endowment"].value)}},
     "numFaculty" : {{jsonize(cells["numFaculty"].value)}},
     "numDoctoral" : {{jsonize(cells["numDoctoral"].value)}},
     "country" : {{jsonize(cells["country"].value)}},
 }
```

图 3.55 “自定义表格导出器”对话框

3.3.12 函数

1．Length(string s)函数

该函数的功能是返回字符串的长度。

例如，查看某列文本的长度，其中，“France”的长度是 6，“USA”的长度是 3，如图 3.56 所示。

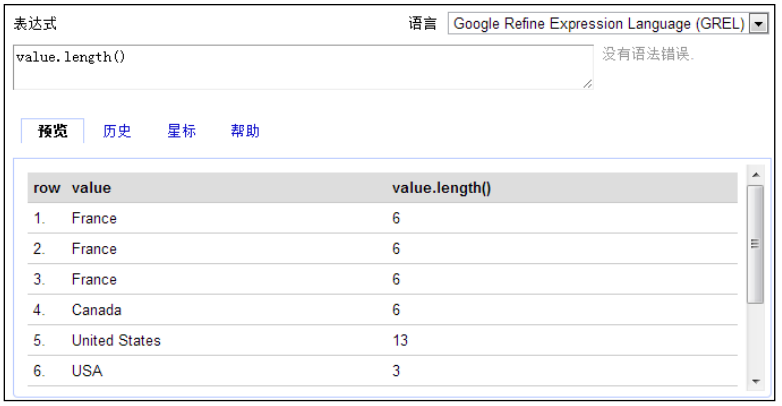


图 3.56　length( )函数用法

2．startsWith(string s, string sub)函数

该函数的功能是判断字符串“s”是否包含前缀字符串“sub”。若包含，则返回“true”，否则返回“false”。

例如，查看某列文本是否以“U”开头，其中“France”字符串不以“U”开头，返回“false”，而字符串“USA”和“United States”均以“U”开头，返回“true”，如图 3.57 所示。注意函数 startsWith( ) 的“W”是大写的。

图 3.57　startsWith( )函数用法

3．endsWith(string s, string sub)函数

该函数的功能是判断字符串“s”是否包含后缀字符串“sub”。若包含，则返回“true”，否则返回“false”。

例如，查看某列文本是否以“a”结尾，其中“Canada”字符串以“a”结尾，返回“true”，而字符串“USA”不以“a”结尾（区分大小写），返回“false”，如图 3.58 所示。

图 3.58　endsWith( )函数用法

4．Contains(string s, string sub)函数

该函数的功能是判断字符串“s”是否包含字符串“sub”。若包含，则返回“true”，否则返回“false”。

例如，查看某列文本是否包含 “n”，其中“France”、“Canada”和 “United States”字符串均包含“n”，返回“true”，而字符串“USA”不包含“n”，返回“false”，如图 3.59 所示。

图 3.59　contains( )函数用法

5．Trim(string s)和 strip(string s)函数

这两个函数功能相同，均返回忽略（去除）开头和结尾的空白的字符串 s 的副本。

如字符串“　保定　”（前后各有三个空格）、“　保定”（前面有两个空格）和“保定　”（后面有两个空格）应用该函数后均返回字符串“保定”（前后均无空格），但字符串中间的空格无法删

除，如“　保　定　”（前后各有三个空格且中间有两个空格）、“　保　定”（前面有两个空格且中间有两个空格）和“保　定　”（后面有两个空格且中间有两个空格）应用该函数后均返回字符串“保　定”（中间有两个空格）。

6．Substring(s, number from, optional number to)函数

该函数的功能是提取字符串“s”中介于“from”和“number to”之间的字符串。注意，OpenRefine中的字符串相当于数组，字符串的下标从“0”开始。

如 substring（"profound", 3）返回字符串“found”，因为下标 3 对应的字符是“f”，没有第二个参数“number to”，则返回字符“f”开始后面的所有字符，即字符串“found”。

如 substring（"France", 2, 4）返回字符串“an”，因为下标 2 对应的字符是“a”，下标 4 对应的字符是“c”，则返回字符“a”开始且字符“c”之前（不含）的字符串，即“an”。

如 substring（"USA", 2, 4）返回字符串“A”，因为下标 2 对应的字符是“A”，下标 4 没有对应的字符，字符串的长度是 3，所以返回字符“A”，如图 3.60 所示。

图 3.60　substring ( )函数用法

7．Replace(string s, string f, string r)函数

该函数的功能是在字符串中用一些字符（或字符串）替换另一些字符（或字符串），或替换一个与正则表达式匹配的子串。如将某列所有“a”替换为“A”，如图 3.61 所示。

replace( )函数格式还可以为“value.replace（"a","A"）. replace（"s","S"），表示将某列所有字符“a”和“s”分别替换为字符“A”和“S”。

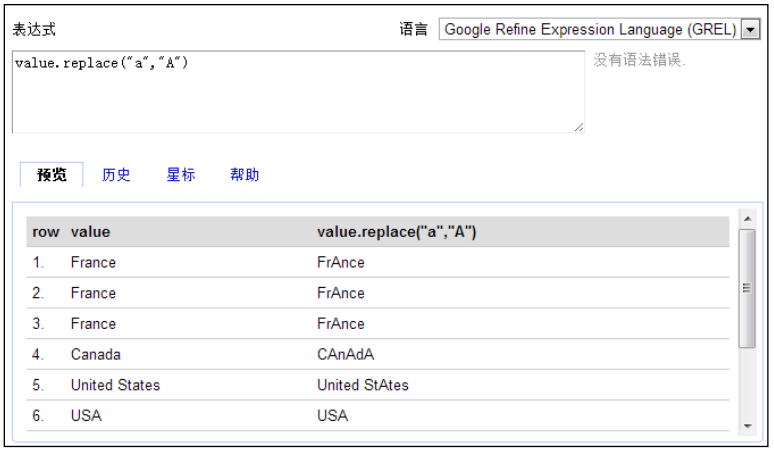


图 3.61　replace( )函数用法

8．Split(s, sep)函数

该函数的功能是返回一个字符串数组，即用 sep 分割字符串 s 得到的数组。

例如，split("Claire Gute"," ")返回数组 ["Claire", "Gute"]。如某列仅想保留数组的一个元素，则该表达式保留某列返回字符串数据下标为零的元素，如图 3.62 所示。若想保留第二个元素，则下标是 1。

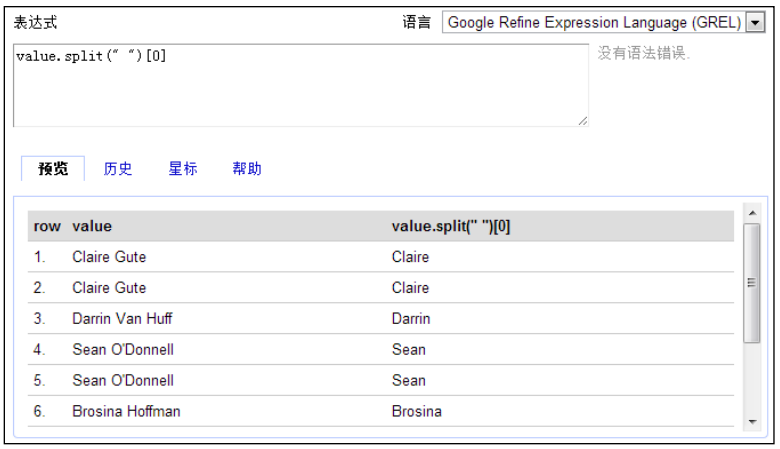


图 3.62　split( )函数用法

9．Min(number d1, number d2)和 max(number d1, number d2)函数

min( )函数的功能是返回数值型 d1 和 d2 中最小的数，max( )函数的功能正好相反。

10．if( )函数

if( )函数的用法与 Excel 中对应的函数用法相同，可参见 3.4.1 小节。如某列对应的属性值大于 10 000，则用原值的 2 倍替换原值，即 value×2，否则用零替代原值，如图 3.63 所示。

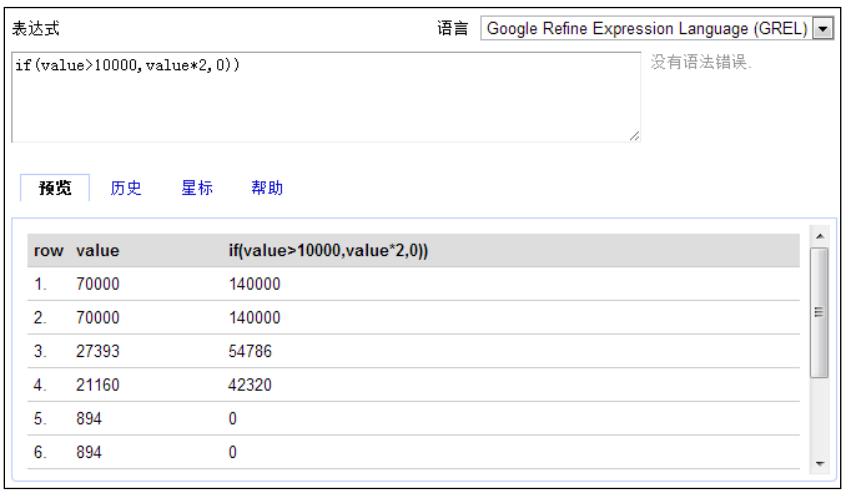


图 3.63　if( )函数用法

3.3.13　正则表达式

正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符或这些特定字符的组合，组成一个“字符串规则”，使用该“字符串规则”对字符串进行过滤、筛选或者从字符串中获取特定的部分。

如图 3.64 所示，使用“文本过滤器”中的正则表达式“[ae]”筛选符合条件的记录，必须勾选“正则表达式”复选框才能实现筛选，“大小写敏感”复选框勾选与否决定正则表达式筛选时是否区分大小写。

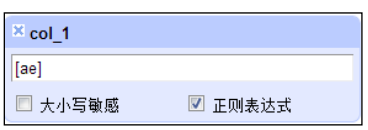


图 3.64　正则表达式用法 1

如图 3.65 所示，使用正则表达式“/.*(?[as])?.*/”匹配某列字段中包含字符“a”或“s”的数组，“[0]”表示显示匹配数组的首个元素。

图 3.65　正则表达式用法 2

如图 3.66 所示，使用正则表达式“/.*（\d{4}）.*/”匹配某列字段中包含连续 4 个数字的数组。

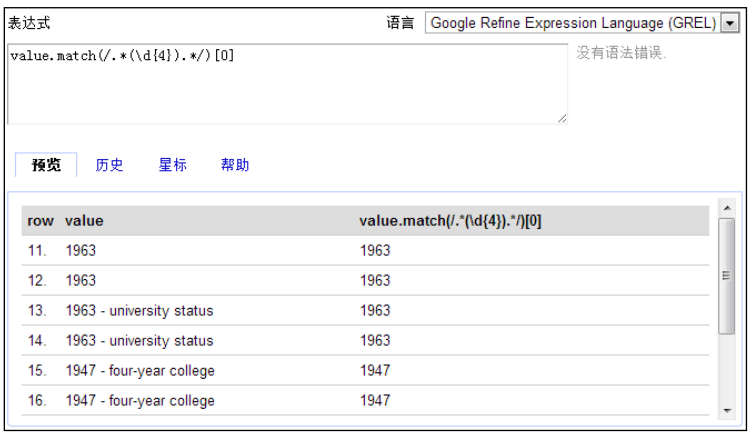


图 3.66　正则表达式用法 3

1．字符类（Character classes）

字符类与一组字符中的任何一个字符匹配，常见的字符类正则表达式如下。

- [character_group]

功能：匹配 character_group 中的任何单个字符，注意区分大小写。

如模式：[ae]，匹配字符串“gray”和“lane”，但不匹配字符串“boss”。

- [^character_group]

功能：与不在 character_group 中的任何单个字符匹配，注意区分大小写。

如模式：[^ae]，匹配字符串“boss”、“gray”和“lane”，而不匹配字符串“aeae”和“aee”。

- [第一个 - 最后一个]

功能：与从第一个到最后一个的范围中的任何单个字符匹配。

如模式：[a-w]，匹配字符串“boss”、“gray”和“lane”，而不匹配字符串“xyz”、“1234”。

- 通配符.

功能：与除 \n（新行）外的任何单个字符匹配。

如模式：a.e，匹配字符串“nave”、“state”，而不匹配字符串“na.ve”、“na?ve”和“lae”，即仅匹配字符“a”和字符“e”之间只包含一个非换行的单个字符。

- \w

功能：匹配一个字母、数字或下划线。

如模式：\w，匹配字符串“nave”、“_??”和“1234”，而不匹配字符串“*??”。

- \W

功能：匹配除字母、数字或下划线外的其他符号。

如模式：\W，匹配字符串“n a v e”、“na.ve ”和“na?ve”，不匹配字符串“nave”和“999”。

- \d

功能：与任何一个十进制数字匹配。

如模式：\d，匹配字符串“123”、“boss123”和“901–”，而不匹配“boss”。

- \D

功能：匹配任何一个非十进制数字。

如模式：\D，匹配字符串“boss”、“boss123”、“–333”和“1234.5”，而不匹配“123”。

思考：

abc、\d[a-z]、[^a-zA-Z0-9]的功能是什么？匹配的字符串具有哪些特征？[1]

2．定位点（Anchors）

断言会使匹配成功或失败，具体取决于字符串中的当前位置，但它们不会使引擎在字符串中前进或使用字符。常见的定位点正则表达式如下。

- ^

功能：匹配必须从字符串的开头开始。

如模式：^\d{3}，匹配字符串“1234”、“999.9”、“901”和“901–”，因为这几个字符串均以 3 个十进制数字开头。不匹配字符串“0.9”、“boss123”、“*??”和“–999”。

1　字符类的含义如下。

abc：匹配包含“abc”的字符串，如“67abcd”，但“adbc”不匹配。

\d[a-z] ：匹配包含一个数字和任意一个字符的字符串，如“abc9abc”和“67abcd”，但“abc89”和“45!ab”不匹配。

[^a-zA-Z0-9]：匹配包含一个非数字非大小写字母的字符串，如“45df　678”、“45!ab”均匹配（包含空格或“!”等特殊符号）。

- $

功能：匹配必须出现在字符串的末尾。

如模式：–\d{3}$，匹配字符串“–333”和“–999”，因为这两个字符串均以 3 个十进制数字结尾，从后面数第四位是“–”。不匹配字符串“–9999”和“–99”。

- \b

功能：匹配必须出现在单词的边界上。

如模式：\bs\w+，匹配字符串“state”、“south Bend”、“state university”，因为字母“s”在单词的边界，后面有一个或一个以上的文字、数字或下划线。不匹配“boss123”、“*??”和“this”。

- \B

功能：匹配不得出现在单词的边界上。

如模式：\Bend\w*\b，匹配字符串“South Bend”、“Rend Lake”，而不匹配字符串“ender”。

思考：

^\d、\d$、^\d.*\d$、\b\d{2}\b、^\d{2}\b的功能是什么？匹配的字符串具有哪些特征？[1]

3．数量词（Quantifiers）

指定在输入字符串中必须存在上一个元素（可以是字符、组或字符类）的多少个实例才能出现匹配项。常见的数量词正则表达式如下。

- *

功能：匹配一个元素零次或多次。

如模式：\d*\.\d，匹配字符串“.9”、“9.9”和“999.9”，不匹配“999.”，即匹配小数点后面至少有 1 位十进制数字、小数点前面有无十进制数字均可的字符串。

- +

功能：匹配一个元素一次或多次。

如模式：be+，匹配字符串“been”和“south bend”中的“be”，即匹配以字符“b”开头，后面至少跟着一个字符“e”的字符串。

- ?

功能：匹配一个元素零次或一次。

如模式：te?，匹配字符串“test”、“bent”、“state”和“tet”，即匹配至少包含一个字母“t”的字符串。

---

1 定位点的含义如下。

^\d：匹配以数字开头的字符串。

\d$：匹配以数字结尾的字符串。

^\d.*\d$：匹配以数字开头并以数字结尾的字符串。

\b\d{2}\b：匹配至少包含一个正好两位数字的字符串（3 位数字的字符串不匹配），如“12”和“–89”。

^\d{2}\b：匹配以两个数字开头的字符串，如“12”，但“–89”不匹配。

- {*n*}

功能：匹配一个元素恰好 *n* 次。

如模式：,\d{3}，匹配字符串“1,234.5”、“,234”和“1,234,567,890”，不匹配字符串“12,34”，即匹配“,”后面恰好有 3 个十进制数字的字符串。

- {*n* ,}

功能：匹配一个元素至少 *n* 次。

如模式：\d{2,}，匹配字符串“123”、“12”和“1234”。

- {*n*,*m*}

功能：匹配一个元素至少 *n* 次，但不多于 *m* 次。

如模式：\d{3,5}，匹配字符串“123”、“12345”和“123456”和“−999”，不匹配字符串“0.9”、“99”。

4．选择（Choices）

用于修改正则表达式以启用“或者”匹配。常见的选择正则表达式如下。

- |

功能：匹配以竖线“|”字符分隔的任何一个元素。

如模式：th（e|is|at），匹配字符串“this”、“the”和“that”。

5．分组（Groups）

分组构造描述了正则表达式的子表达式，通常用于捕获输入字符串的子字符串。常见的分组正则表达式如下。

- （子表达式）

功能：精确匹配括号中的完整子表达式。

如模式：（ae）+，匹配字符串“ae”、“aeae”和“aeee”。

3.4 使用 Excel 简单分析数据

3.4.1　常用函数

Excel 包含的函数其实是一些预定义的公式，函数包含函数名、括号及括号中的参数，参数是一些特定数值，可以按特定的顺序或结构进行计算。

如图 3.67 所示的数据来源于 2015 年 12 月 20 日网站 http://www.cnemc.cn/ 首页的数据，图中仅使用了前 14 条记录。

	A	B	C	D
1	地区	首要污染物	等级	AQI
2	北京市	PM2.5	中度污染	168
3	天津市	NO2	良	87
4	石家庄市	PM2.5	轻度污染	140
5	唐山市	PM2.5	轻度污染	110
6	秦皇岛市	NO2	良	92
7	邯郸市	PM2.5	轻度污染	138
8	邢台市	PM2.5	重度污染	209
9	保定市	PM2.5	重度污染	269
10	承德市	PM2.5	良	94
11	沧州市	PM2.5	中度污染	153
12	廊坊市	PM2.5	中度污染	170
13	衡水市	PM2.5	重度污染	268
14	张家口市	PM10	良	77
15	太原市	PM2.5	轻度污染	129

图 3.67　常用函数的原始数据

1．COUNT( )函数

该函数计算参数列表中的数字项的个数。函数 COUNT( )在计数时，将把数值型数据计算进去，而错误值、空值、逻辑值和文字则被忽略。该函数的语法规则如下：

COUNT（value1,value2,…）

参数：value1,value2,…是包含或引用各种类型数据的参数（1～30 个）。

例如：“=COUNT（B2:B15）”的值是 0，因为数据区域“B2:B15”中没有数值型的数据。

“=COUNT（D2:D15）”的值是 14，因为数据区域“D2:D15”中共有 14 个数值型数据。

2．COUNTIF( )函数

对指定区域中符合指定条件的单元格计数。该函数的语法规则如下：

COUNTIF（range，criteria）

参数：range 是计算其中非空单元格数目的区域。

criteria 是数字、表达式或文本形式定义的条件。

例如：“=COUNTIF（C2:C15,"重度污染"）”的值是 3。

3．SUM( )和 AVERAGE( )函数

返回某一个单元格区域中数字、逻辑值及数字的文本表达式之和或均值。如果参数中有错误值或不能转换成数字的文本，将会导致错误。这两个函数的语法规则如下：

SUM（number1,number2,…）

AVERAGE（number1,number2,…）

参数：number1,number2,…是对其求和或均值的 1 到 255 个参数。

例如：“=SUM（D2:D15）”的值是 2104。

“=AVERAGE （D2:D15）”的值是 150.3。

4．SUMIF( )和 AVERAGEIF( )函数

SUMIF( )和 AVERAGEIF( )函数按给定条件对指定单元格求和或求平均值。这两个函数的语法规则如下：

SUMIF（range, criteria, sum_range）

AVERAGEIF（range, criteria, average_range）

参数：range 是根据条件计算的单元格区域。每个区域中的单元格都必须是数字和名称、数组和包含数字的引用。空值和文本值将被忽略。

criteria 是确定对哪些单元格相加的条件，其形式可以为数字、表达式或文本。例如，条件可以表示为 32、"32"、">32" 或 "apples"。

sum_range/ average_range 为要相加或求均值的实际单元格（如果区域内的相关单元格符合条件）。

例如：“=SUMIF（C2:C15,"重度污染",D2:D15）”的值是 746。

“=AVERAGEIF（C2:C15,"重度污染",D2:D15）”的值是 248.7。

5．MAX( )和 MIN( )函数

返回一组值中的最大值或最小值。这两个函数的语法规则如下：

MAX（number1,number2,…）

MIN（number1,number2,…）

参数：number1,number2,…是要从中找出最大值或最小值的 1 到 255 个数字参数。

例如：“=MAX（D2:D15）”的值是 269。

“=Min（D2:D15）”的值是 77。

6．RANK( )函数

RANK( )函数返回一个数字在数字列表中的排位。数字的排位是其大小与列表中其他值的比值（如果列表已排过序，则数字的排位就是它当前的位置）。该函数的语法规则如下：

RANK（number, ref, order）

参数：number 是需要排位的数字。

ref 是数字列表数组或对数字列表的引用，非数值型参数将被忽略。

order 是一个数字，指明排位的方式。如果 order 为 0（零）或省略，则 Excel 对数字的排位是基于 ref 按照降序排列的列表。如果 order 不为零，则 Excel 对数字的排位是基于 ref 按照升序排列的列表。

例如：在 E2 单元格中输入“=RANK（D2,$D$2:$D$15,1）”并复制到 E 列其他单元格，在 F2 单元格中输入“=RANK（D2,$D$2:$D$15,0）”并复制到 F 列其他单元格，结果如图 3.68 所示。

	A	B	C	D	E	F
1	地区	首要污染物	等级	AQI	升序	降序
2	北京市	PM2.5	中度污染	168	10	5
3	天津市	NO2	良	87	2	13
4	石家庄市	PM2.5	轻度污染	140	8	7
5	唐山市	PM2.5	轻度污染	110	5	10
6	秦皇岛市	NO2	良	92	3	12
7	邯郸市	PM2.5	轻度污染	138	7	8
8	邢台市	PM2.5	重度污染	209	12	3
9	保定市	PM2.5	重度污染	269	14	1
10	承德市	PM2.5	良	94	4	11
11	沧州市	PM2.5	中度污染	153	9	6
12	廊坊市	PM2.5	中度污染	170	11	4
13	衡水市	PM2.5	重度污染	268	13	2
14	张家口市	PM10	良	77	1	14
15	太原市	PM2.5	轻度污染	129	6	9

图 3.68　RANK( )函数的使用

7．IF( )函数

IF( )函数根据指定的条件来判断其“真”或“假”并返回“TRUE”或“FALSE”值。该函数的语法规则如下：

IF（logical_test,value_if_true,value_if_false）

参数：logical_test 是判断条件。

value_if_true 是为“TRUE”时返回的值。

value_if_false 是为“FALSE”时返回的值。

例如：“=IF（D2>150,"可以户外活动","不建议室外活动"）”的值是“可以户外活动”。

8．TRIM( )函数

此函数的用法与 OpenRefine 对应函数的用法相同，可参见 3.3.12 小节。

3.4.2　筛选

筛选操作仅显示满足指定条件的记录，不满足条件的记录被隐藏。

将活动单元格放在数据表中，然后在“开始”选项卡上的“编辑”组中，单击“排序和筛选”，然后单击“筛选”，再单击列标题中的箭头按钮实现筛选操作。

在文本值列表中，选择或清除一个或多个要作为筛选依据的文本值，也可以指定“文本筛选”，设置筛选条件，其菜单如图 3.69 所示。

在数值列表中，选择或清除一个或多个要作为筛选依据的数值，也可以指定“数字筛选”，设置筛选条件，其菜单如图 3.70 所示。

图 3.69　“文本筛选”菜单

图 3.70　“数字筛选”菜单

如筛选“等级”中以“污染”结尾的数据。首先实现筛选操作，然后在“等级”列中单击三角形下拉按钮，选择【文本筛选】|【结尾是】选项，在打开的对话框中输入“污染”，如图 3.71 所示，单击“确定”按钮后在“等级”列中筛选出以“污染”结尾的记录，如图 3.72 所示。注意“等级”列的筛选图标，单击该图标后选择“从‘等级’中清除筛选”可以清除这个筛选操作。

在“开始”选项卡的“编辑”组中，单击“排序和筛选”，再单击“筛选”，可以取消筛选操作。

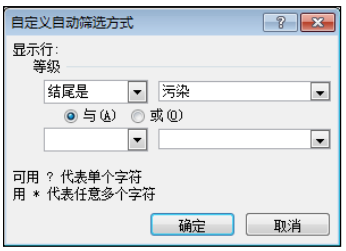


图 3.71　文本“结尾是”筛选

	A	B	C
1	地区	首要污染物	等级
2	保定市	PM2.5	重度污染
3	北京市	PM2.5	中度污染
4	沧州市	PM2.5	中度污染
6	邯郸市	PM2.5	轻度污染
7	衡水市	PM2.5	重度污染
8	廊坊市	PM2.5	中度污染
10	石家庄市	PM2.5	轻度污染
11	太原市	PM2.5	轻度污染
12	唐山市	PM2.5	轻度污染
14	邢台市	PM2.5	重度污染

图 3.72　文本筛选结果

3.4.3　数据透视表（PivotTable）

使用函数和筛选功能可以对数据进行简单的统计，当记录较多时，还可以使用 Excel 数据透视表（PivotTable）快速汇总大量数据，深入分析数据，并且可以回答一些预计不到的数据问题。

首先选择要进行统计的工作表数据，然后单击“插入”选项卡中的“数据透视表”即可创建数据透视表。如根据图 3.67 所示的原始数据制作一个数据透视表。在打开的对话框中设置要分析的数据区域，然后选择放置数据透视表的位置，如图 3.73 所示。单击“确定”按钮后设置数据透视表字段列表，行标签是“首要污染物”，列标签是“等级”，数值是“计数项：地区”，如图 3.74 所示。创建好的数据透视表如图 3.75 所示。

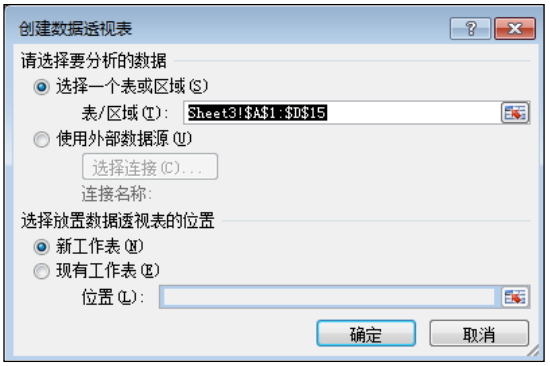


图 3.73　创建数据透视表

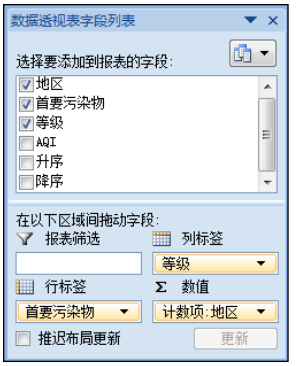


图 3.74　数据透视表字段列表

计数项:地区	列标签				
行标签	良	轻度污染	中度污染	重度污染	总计
NO2	2				2
PM10	1				1
PM2.5	1	4	3	3	11
总计	4	4	3	3	14

图 3.75　创建好的数据透视表

从数据透视表可以直观地看出共有 14 个地区，从行上看，其中 2 个地区的首要污染物是“NO2”，1 个地区的首要污染物是“PM10”，11 个地区的首要污染物是“PM2.5”。从列上看，等级为“良”的地区有 4 个，等级为“轻度污染”的地区有 4 个，等级为“重度污染”的地区有 3 个，等级为“中度污染”的地区有 3 个。从行列同时看，可以看出首要污染物是“PM2.5”且等级为“中度污染”的地区是 3 个。

在数据透视表的“设计”视图中可以修改数据表的行和列，如图 3.76 所示是行列互换后的数据透视表。

	A	B	C	D	E
1	计数项:地区	列标签			
2	行标签	NO2	PM10	PM2.5	总计
3	良	2	1	1	4
4	轻度污染			4	4
5	中度污染			3	3
6	重度污染			3	3
7	总计	2	1	11	14

图 3.76　行列互换后的数据透视表

3.4.4　在透视表里做筛选

在数据透视表中可以再做筛选，如单击“行标签”或“列标签”的三角形下拉按钮，可以进行“标签筛选”、“值筛选”，也可以勾选筛选记录，如图 3.77 所示。

将鼠标指针放在数据透视表“列”的右侧一列单元格中，如“F2”单元格，在“开始”选项卡的“编辑”组中，单击“排序和筛选”，然后单击“筛选”，也可以实现透视表中的筛选，如图 3.78 所示。

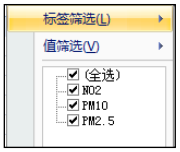


图 3.77　透视表中的筛选 1

	A	B	C	D	E
1	计数项:地区	列标签			
2	行标签	NO2	PM1	PM2.	总计
3	良	2	1	1	4
4	轻度污染			4	4
5	中度污染			3	3
6	重度污染			3	3
7	总计	2	1	11	14

图 3.78　透视表中的筛选 2

如筛选首要污染物是“PM2.5”且等级为“中度污染”和“重度污染”不少于 2 的地区个数，筛选结果如图 3.79 所示。注意记录号是“3”和“4”的记录不符合筛选条件，因此隐藏显示。

	A	B	C	D	E
1	计数项:地区	列标签			
2	行标签	NO2	PM1	PM2.	总计
5	中度污染			3	3
6	重度污染			3	3
7	总计	2	1	11	14

图 3.79　透视表中的筛选 3

3.5 数据清理原则

数据清理过程中经常会遇到这样或那样的问题，掌握一些原则和技巧可以让工作事半功倍。

备份原文件。无论数据量是大是小，保存为哪种格式，一定要备份原文件。大量的清理工作都应该在备份文件上操作，有任何数据上的问题方便查询原文件，因为清理工作往往不是一个人短期内可以完成的，在不同的阶段往往由不同的人使用不同的清理方法，甚至使用不同的清理工具。

抽查数据的重要性。在每个阶段都要养成抽查数据的习惯，如每次编辑单元格后都要随机抽查 5 ~ 10 条记录，尽早发现问题。有时候，很多问题看上去很简单，如 3.6.1 小节中的“查找重复记录”案例，但这会影响归类的结果，对后期的数据分析和诠释数据都有非常大的影响。

关注文本数据。文本数据要特别注意空格、大小写和对齐方式等，还要特别关注函数，如果文本函数的结果异常，很可能是因为文本包含乱码。

测试数据。在对数据进行编辑前，要仔细查看“预览”选项卡，否则单击“确定”按钮后，所有记录都将发生改变(可能数据是 5 万条或 10 万条), 尽可能在进行这类操作前确认操作的正确性。

记录清理过程。OpenRefine 导出项目时会自动记录清理过程，但 Excel 没有这项功能，要尽可能地记录所有清理过程，因为清理中撤销某个错误步骤是难免的。

3.6 综合案例

3.6.1 查找重复记录

要求：查找如图 3.80 所示的“地区”列中重复的城市记录。

	A	B	C	D
1	地区	首要污染物	等级	日期
2	保定	PM2.5	重度污染	2015/12/20
3	保定	PM2.5	重度污染	2015/12/21
4	保定	PM2.6	重度污染	2015/12/22
5	保定	PM2.5	重度污染	2015/12/23
6	北京	PM2.5	中度污染	2015/12/20
7	北 京	PM2.6	重度污染	2015/12/21
8	北京	PM2.7	重度污染	2015/12/22
9	北京	PM2.8	重度污染	2015/12/23

图 3.80　查找重复记录的原始数据

方法一：使用 COUNTIF( )函数

在 E2 单元格中输入函数“=IF（COUNTIF（A:A,A2）>1,"重复","")”，然后复制到 E3:E9，结果如图 3.81 所示。

	A	B	C	D	E
1	地区	首要污染物	等级	日期	重复否
2	保定	PM2.5	重度污染	2015/12/20	重复
3	保定	PM2.5	重度污染	2015/12/21	
4	保定	PM2.6	重度污染	2015/12/22	
5	保定	PM2.5	重度污染	2015/12/23	重复
6	北京	PM2.5	中度污染	2015/12/20	重复
7	北 京	PM2.6	重度污染	2015/12/21	
8	北京	PM2.7	重度污染	2015/12/22	重复
9	北京	PM2.8	重度污染	2015/12/23	重复

图 3.81　使用 COUNTIF( )函数查找重复记录

A2:A5 区域的单元格内容似乎相同，尤其是 A2 和 A3 单元格，看起来完全相同，为什么函数计算后不是“重复”呢？原因是 A3 单元格的实际内容是“保定 ”（注意后面有空格），因为空格无法显示出来，所以 A2 和 A3 单元格看起来完全相同，但 Excel 知道二者是完全不同的，A4 单元格的实际内容是“ 保定”（注意前面有空格）。同样，A7 单元格的实际内容是“北 京”（注意中间有空格），与 A8 和 A9 单元格也不相同。

去掉文本前后空格的方式是使用函数 TRIM( )，其功能是除了文本中间的空格外，清除文本前后的所有空格。该函数的语法规则是 TRIM（text），其中参数 text 是需要清除前后空格的文本。

在 F2 单元格中输入函数“=IF（COUNTIF（A:A,TRIM（A2））>1,"重复","")”，然后复制到 F3:F9，结果如图 3.82 所示。

	A	B	C	D	E	F
1	地区	首要污染物	等级	日期	重复否	重复否(trim函数)
2	保定	PM2.5	重度污染	2015/12/20	重复	重复
3	保定	PM2.5	重度污染	2015/12/21		重复
4	保定	PM2.6	重度污染	2015/12/22		重复
5	保定	PM2.5	重度污染	2015/12/23	重复	重复
6	北京	PM2.5	中度污染	2015/12/20	重复	重复
7	北 京	PM2.6	重度污染	2015/12/21		
8	北京	PM2.7	重度污染	2015/12/22	重复	重复
9	北京	PM2.8	重度污染	2015/12/23	重复	重复

图 3.82　使用 COUNTIF( )和 TRIM( )函数查找重复记录

TRIM( )函数无法删除文本中间的空格，如 A7 单元格的内容是“北　京”(注意中间有空格)，即使使用 TRIM( )函数，A7 单元格的内容依旧是“北　京”。

方法二：非公式法

选中需要进行数据重复检查的列或区域 A2:A9，然后单击“开始”选项卡，选择“样式”选项组中的“条件格式”选项，在下拉列表的第一个选项“突出显示单元格规则”中选择“重复值”，在打开的“重复值”对话框中设置“重复”值为“浅红填充色深红色文本”即可，结果如图 3.83 所示。

	A	B	C	D
1	地区	首要污染物	等级	日期
2	保定	PM2.5	重度污染	2015/12/20
3	保定	PM2.5	重度污染	2015/12/21
4	保定	PM2.6	重度污染	2015/12/22
5	保定	PM2.5	重度污染	2015/12/23
6	北京	PM2.5	中度污染	2015/12/20
7	北　京	PM2.6	重度污染	2015/12/21
8	北京	PM2.7	重度污染	2015/12/22
9	北京	PM2.8	重度污染	2015/12/23

图 3.83　使用“突出显示单元格规则”查找重复记录

方法三：数据透视表

选中区域 A1:D9，然后单击“插入”选项卡，选择“数据透视表”，设置“数据透视表字段列表”，如图 3.84 所示。将行标签设置为“地区”，将数值设置为“计数项：地区”，结果如图 3.85 所示。

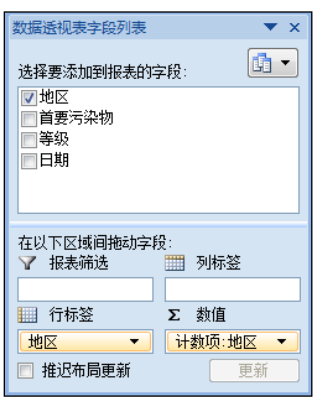


图 3.84　设置“数据透视表字段列表”

行标签	计数项:地区
保定	1
保定	2
保定	1
北　京	1
北京	3
总计	8

图 3.85　使用“数据透视表”查找重复记录

通过图 3.85 可以看出，“ 保定”(前面有空格) 1 项，“保定”(前后均无空格) 2 项，“保定 ”(后面有空格) 1 项。

特别注意，用分类汇总是不能查找重复记录的。通过图 3.86 可以看出，以“地区”为分类字段汇总“地区”计数时，不区分“ 保定”(前面有空格)、“保定”(前后均无空格) 和“保定 ”(后面有空格)，认为三者是一样的。但区分“北　京”(中间有空格) 和“北京”(中间无空格)，结果如图 3.87 所示。

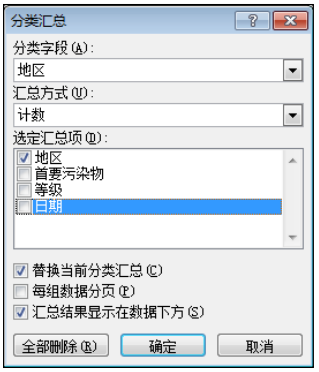


图 3.86　设置“分类汇总”

	A	B	C	D	E
1		地区	首要污染物	等级	日期
2		保定	PM2.5	重度污染	2015/12/20
3		保定	PM2.5	重度污染	2015/12/21
4		保定	PM2.6	重度污染	2015/12/22
5		保定	PM2.5	重度污染	2015/12/23
6	保定 计数	4			
7		北京	PM2.5	中度污染	2015/12/20
8	北京 计数	1			
9		北 京	PM2.6	重度污染	2015/12/21
10	北 京 计数	1			
11		北京	PM2.7	重度污染	2015/12/22
12		北京	PM2.8	重度污染	2015/12/23
13	北京 计数	2			
14	总计数	8			

图 3.87　使用“分类汇总”查找重复记录（不正确）

使用 OpenRefine 对“地区”列进行“文本归类”，如图 3.88 所示。单击“簇集”按钮将“ 保定”（前面有空格）和“保定 ”（后面有空格）合并为“保定”，如图 3.89 所示。

图 3.88　对“地区”列进行“文本归类”

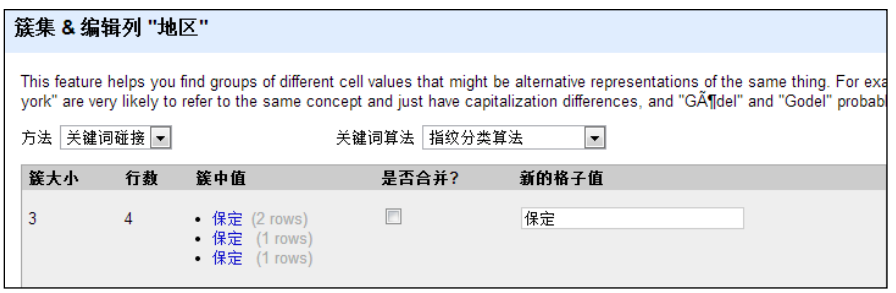


图 3.89　合并簇集

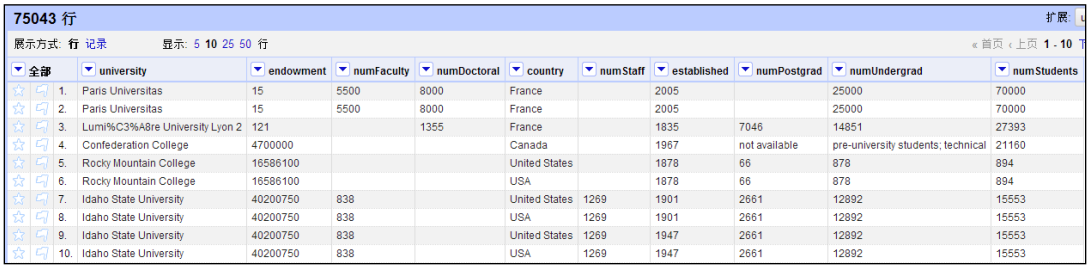
3.6.2　使用 OpenRefine 清理数据

本案例使用 OpenRefine 对数据集进行数据清理。首先从站点下载源数据压缩包“http://enipedia.tudelft.nl/enipedia/images/f/ff/UniversityData.zip”，解压为文件“UniversityData.csv”。

（1）导入数据。

首先运行 OpenRefine，然后单击“新建项目”，“数据来源于”选择“这台电脑”，可以选择存储

在本地计算机上的数据文件，导入文件“UniversityData.csv”，在主界面查看数据，如图 3.90 所示。



75043 行　　扩展:

展示方式: 行 记录　　显示: 5 10 25 50 行　　« 首页 ‹ 上页 1 - 10 下

全部	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
1.	Paris Universitas	15	5500	8000	France		2005		25000	70000
2.	Paris Universitas	15	5500	8000	France		2005		25000	70000
3.	Lumi%C3%A8re University Lyon 2	121		1355	France		1835	7046	14851	27393
4.	Confederation College	4700000			Canada		1967	not available	pre-university students; technical	21160
5.	Rocky Mountain College	16586100			United States		1878	66	878	894
6.	Rocky Mountain College	16586100			USA		1878	66	878	894
7.	Idaho State University	40200750	838		United States	1269	1901	2661	12892	15553
8.	Idaho State University	40200750	838		USA	1269	1901	2661	12892	15553
9.	Idaho State University	40200750	838		United States	1269	1947	2661	12892	15553
10.	Idaho State University	40200750	838		USA	1269	1947	2661	12892	15553

图 3.90　查看 UniversityData 数据

（2）发现问题。查看主界面，发现数据存在以下问题。

- 有空值记录。如第 3 条记录“numFaculty”列的值为空。
- 记录值不正确。如第 4 条记录“numPostgrad”列的值是“not available”，“numUndergrad”列的值是“pre-university students; technical”，正确值应该是数字。
- 所有列的数据类型均为文本型，如“numStudents”列用于记录学生人数，应该是数值型数据（如果在导入数据时“配置解析选项”部分勾选“将单元格中的文本解析为数字、日期……”，则可以正确解析数字和日期等）。
- 记录值有乱码。如第 3 条记录 “university”列的值是“Lumi%C3%A8re University Lyon 2”。

（3）使用“文本归类”清理数据。

- 对 “country” 列进行“文本归类”，发现数据录入错误，有 2 条记录“country”列的值是“,”，如图 3.91 所示。
- 数据录入错误，有 576 条记录“country”列的值是“Canada B1P 6L2”，如图 3.92 所示，正确值应该是“Canada”。

country　修改

68 choices 排序，按照: 名称 数量　簇集

, 2

Albania 8

Argentina 1

图 3.91　录入错误 1

country　修改

68 choices 排序，按照: 名称 数量　簇集

Canada B1P 6L2 576

Canada C1A 4P3 Telephone: 902-566-0439 Fax: 902-566-0795 1

Chile 1

图 3.92　录入错误 2

- 同一个国家有多种录入值。如英国有三种不同的表示方法，其中 3398 条记录用的是“England”，324 条记录用的是“England, UK”，572 条记录用的是“England, United Kingdom”，如图 3.93 所示。
- 有重复记录。如按“university”列文本归类，如图 3.94 所示。选择“Indian Institute of Technology Delhi”，包含 2 条记录，如图 3.95 所示，这两条记录完全相同。

图 3.93　录入错误 3

图 3.94　文本归类

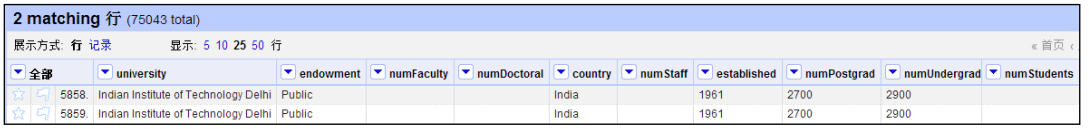


图 3.95　重复记录

随着数据清理的深入，我们还会发现其他问题，下面首先来解决已经发现的问题。

（4）修改格式。

单击“全部移除”按钮移除所有归类。如果记录没有正确解析格式，如“numStudents”列为文本，则选择该列，单击【编辑单元格】|【常用转换】|【数字化】选项将该列转换为数值型数据，同样将列“endowment”、“numFaculty”、“numDoctoral”、“numStaff”、“established”、“numPostgrad”和“numUndergrad”数字化。

（5）修改“country”列国家名字不正确、同国不同名的问题。

单击“全部移除”按钮移除所有归类。选择“country”列，单击【归类】|【文本归类】选项，结果如图 3.96 所示。

查看第一个文本分类“,”，没有国家的名字是“,”，这个国家名字一定是错误的，如何编辑该国家名字呢？查看这 2 条记录的“university”列，显示列值是“Universidad Ju%C3%A1rez Aut%C3%B3noma de Tabasco”，其中包括“%C3%”等特殊字符，初步分析应该是国家的文字乱码导致的。选择“numStudents”列，单击【编辑单元格】|【转换】选项，输入表达式“value.unescape('url')”，将乱码转换为“Universidad Juárez Autónoma de Tabasco”。也可以使用搜索引擎搜索该文本，发现该校名应该是“Universidad Juárez Autónoma de Tabasco”，深入搜索该大学是墨西哥的一所大学 [1]。选择文本分类“,”，单击“编辑”按钮，输入“Mexico”，再单击“应用”按钮，则分类变为 67 个，如图 3.97 所示。

图 3.96　68 个文本归类

图 3.97　67 个文本归类

1 https://en.wikipedia.org/wiki/Universidad_Ju%C3%A1rez_Aut%C3%B3noma_de_Tabasco。

查看包含 2 条记录的“Cura%C3%A7ao”分类和包含 1 条记录的“Nassau, Bahamas Fort Myers, FL Jacksonville, FL Miami, FL Miramar, FL Orlando, FL Palm Beach, FL Tampa, FL”分类，如图 3.98 和图 3.99 所示，尝试用上述方法编辑该值。

图 3.98　“Cura%C3%A7ao”分类

图 3.99　“Nassau, Bahamas……”分类

查看“Canada”附近的分类，如图 3.100 所示。发现 “Canada B1P 6L2”和“Canada C1A 4P3 Telephone: 902-566-0439 Fax: 902-566-0795”均应该是“Canada”，编辑这两个分类，则分类变为 65 个，如图 3.101 所示。

图 3.100　“Canada”文本归类

图 3.101　65 个文本归类

单击“country”分类右上方的“全部重置”按钮清空所有归类，然后单击“簇集”按钮，在“簇集&编辑列‘country’”窗口中查看到共有 3 个簇集，勾选“是否合并”复选框，输入“新的格子值”是“USA”，如图 3.102 所示。这时分类变为 60 个。

图 3.102　簇集

用类似的方法，修改英国的多种不同写法，如图 3.103 所示。修改后分类变为 58 个。

即使这样修改，可能还存在同国不同名的问题，如图 3.104 所示，可以发现还存在英国和美国的其他写法，也存在“Utopia”这种并不存在的国家名。将“United Kingdom”编辑为将“UK”，将“United States of America”和“Utopia”均编辑为“USA”。修改后分类变为 55 个。

图 3.103　修改英国的不同写法

图 3.104　仍然有不同写法存在

用同样的方法，将“Republic of China”编辑为“China”，如图 3.105 所示。修改后分类变为 54 个。将“Rossija”和“Russian Federation”编辑为“Russia”，如图 3.106 所示。修改后分类变为 52 个。

图 3.105　修改中国的不同写法

图 3.106　修改俄罗斯的不同写法

用同样的方法，将“Scotland”、“Scotland, UK”和“Scotland, United Kingdom”均编辑为“UK”，如图 3.107 所示。修改后分类变为 49 个。将“the Netherlands”编辑为“Netherlands”，如图 3.108 所示。修改后分类变为 48 个。

图 3.107　修改英国的不同写法

图 3.108　修改荷兰的不同写法

查看“Satellite locations:”分类，使用搜索工具确定其为美国的一所大学，所以编辑为“USA”，这时分类变为 47 个。将“Taiwan”编辑为“China”，修改后分类变为 46 个。

思考：

查看“country”列，尝试查找是否存在其他需要修改的国家名。

（6）修改“numStudents”列中记录值不正确的问题。

单击“全部移除”按钮移除所有归类。选择“numStudents”列，单击【归类】|【数值归类】选项，结果如图 3.109 所示。显示共有 4695 条非数值型记录和 19269 条空记录，这将是本步骤着重解决的问题。仅勾选“Non-numeric”复选框，查看非数值型记录，如图 3.110 所示。

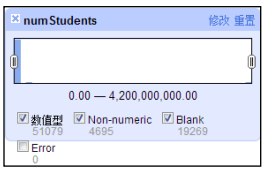


图 3.109　“numStudents”列数值归类

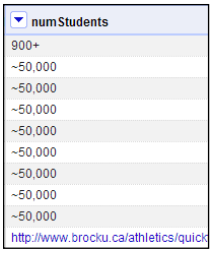


图 3.110　非数值型记录

非数值型记录是左对齐显示的，并且是黑色显示（数值型数据是右对齐，绿色显示）。单击“下一页”按钮查看这类记录的特征。这类记录中很多包含了数字和文字两部分，如“900+”表示 900 余名学生，“~50,000”表示大约 50 000 名学生。

选择“numStudents”列，单击【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“value.replace（"+",""）.replace（"~",""）.replace（",",""）”，表示将字符“+”、“~”和“,”转换为空，即删除这些字符，如图 3.111 所示。

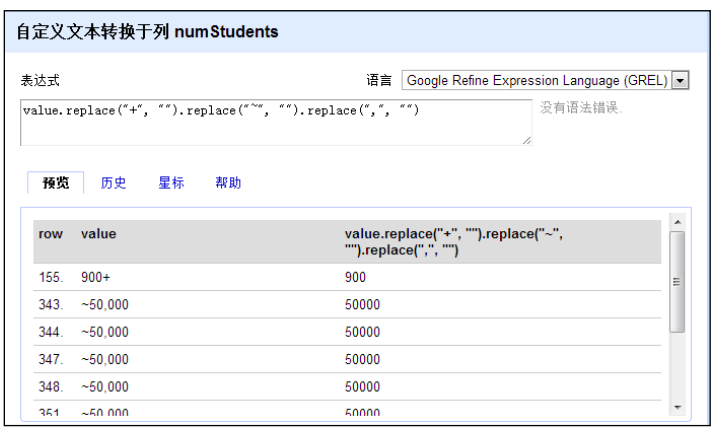


图 3.111　使用 replace( )函数转换单元格

选择“numStudents”列，单击【编辑单元格】|【常用转换】|【数字化】选项，将该列转换为数值型数据。这时显示非数值型记录共有 4678 条。

用同样的方法输入表达式“value.replace（"total",""）.replace（"-",""）”，删除字符串“total”和“-”，然后数字化该列，显示非数值型记录共有 4510 条。

查看剩下的 4510 条记录，很多记录的特征是前面是数字、后面是文字，为解决这个问题，选择“numStudents”列，单击【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“substring（value,0,indexOf（value," "））”，如图 3.112 所示。这个表达式首先通过 indexOf( )函数定位首个空格的下标，然后使用 substring( )函数从开始到下标位置取子字符串，即取出第一个空格之前的数字部分。然后数字化该列，显示非数值型记录共有 30 条。

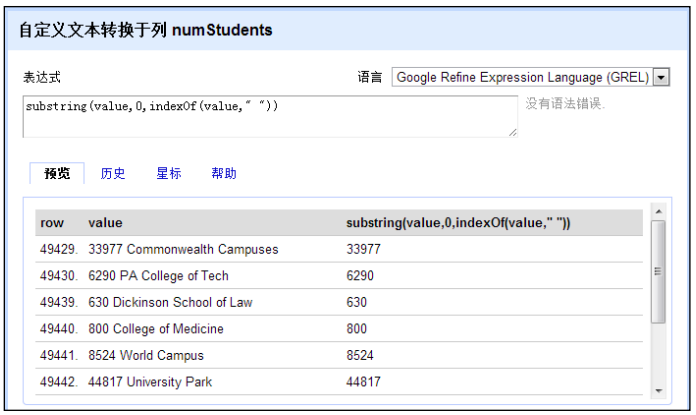


图 3.112　substring( )和 indexOf( )函数的使用

剩下的 30 条记录不值得手工操作修改（相对 75043 条记录来说，30 条记录所占比例非常小），可以直接删除这些记录。

单击【全部】|【编辑行】|【移除所有匹配的行】选项删除这些记录。单击“全部重置”按钮显示所有的记录，更新后的记录共 75 013 行。

同上述方法删除所有空值记录（见图 3.109，空记录共 19 269 条），更新后的记录共 55 744 行。

思考：

尝试使用上述方法编辑“numUndergrad”列的其他数值型数据。

（7）“endowment”列单位不统一的问题。

单击“全部移除”按钮移除所有归类。选择“endowment”列，单击【归类】|【数值归类】选项，结果如图 3.113 所示。显示共有 21 591 条非数值型记录，这将是本步骤着重解决的问题。仅勾选“Non-numeric”复选框，查看非数值型记录，如图 3.114 所示。

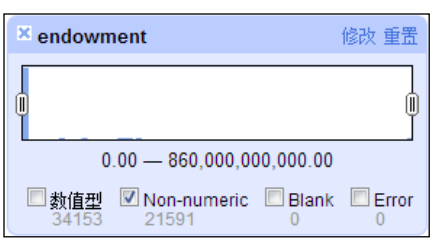


图 3.113　“endowment”列数值归类

图 3.114　非数值型记录

“endowment”列中的数值大部分是以美元表示的，但美元的表示方法很多，如“US $”、“US$”、“$”和“USD $”等，选择“endowment”列，单击【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“value.replace ("US $","") .replace ("US$", "") .replace ("$","") .replace ("USD $","") .replace (",","")”。然后数字化该列，显示非数值型记录共有 21 202 条。

接下来将单位“million”转换为相应的数据。为避免大小写敏感的问题，首先将数据全部转换为小写。选择“endowment”列，单击【编辑单元格】|【常用转换】|【全部小写】选项，然后单击【归类】|【自定义文本归类】选项，在打开的对话框中输入表达式“value.contains ("million")”，如图 3.115 所示。

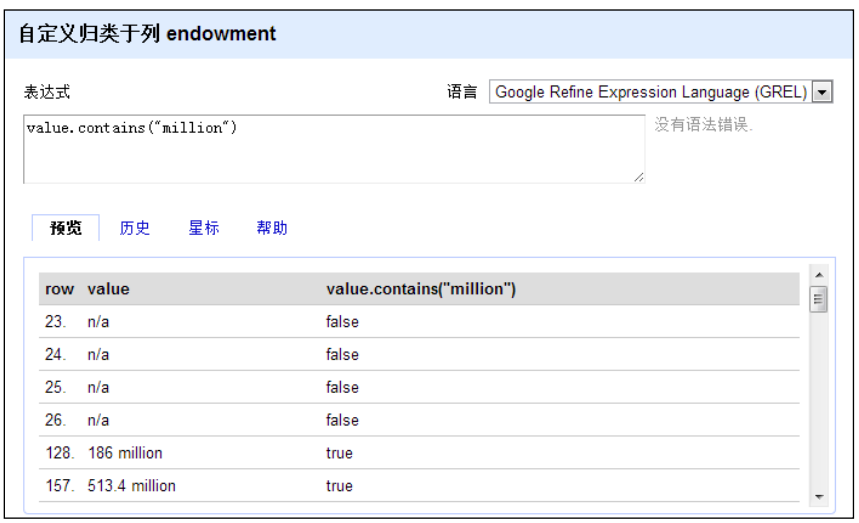


图 3.115　自定义文本归类

记录归为两类，一类包含字符串“million”，共 1381 条记录；另一类不包含“million”字符串，共 19 821 条记录，如图 3.116 所示。选择“true”类，在“endowment”列中单击【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“toNumber (value.replace ("million","")) *1000000”，如图 3.117 所示。然后数字化该列，显示非数值型记录共有 283 条（注意 283 条记录是包含字符串“million”的非数值型数据，实际上还有很多不包含“million”的非数值型数据）。

图 3.116　自定义文本归类结果

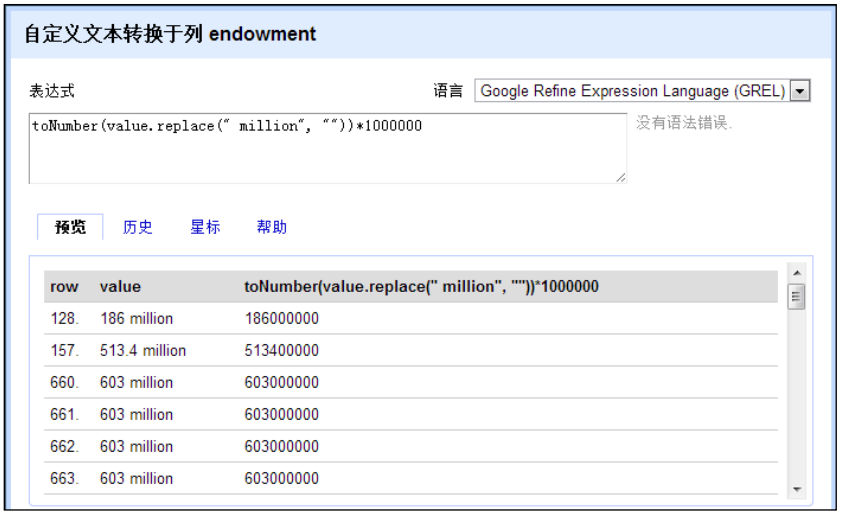


图 3.117　编辑字符串“million”

选择“endowment”列，单击【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“value.replace("r","").replace("u.s.","").replace("p","").replace("+","").replace("usd","").replace ("approx.","")”。然后用同样的方法转换字符串“million”为“1000000”，并数字化列。

思考：

尝试继续清理包含字符串“million”的非数值型数据。

“endowment”列中有些记录以字符“a”开头表示澳元，以字符“c”开头表示加元，尝试按当日汇率转换为美元。

尝试使用上述方法将单位“billion” 转换为相应的数据。

（8）“established”列包含非数值型数据的问题。

单击“全部移除”按钮移除所有归类。“established”列应该显示大学的创建时间，观察记录发现，该列日期混乱，大部分是年份，还有些是文本型日期，或者是文本开头后面是年份等。为了更深入地发现问题，数字化该列后对该列做数值归类，如图 3.118 所示。仅勾选“Non-numeric”复选框查看非数值型数据，如图 3.119 所示。大部分的记录仅包含确切的年份，本列虽然应该是日期型数据，但从记录值上看保存为 4 位年份的数值型数据更为合适，但年份有些保存在字符串开始部分，有些保存在字符串结尾，还有些保存在字符串中间，不适合使用步骤 6 中的 indexOf( )函数和 substring( )函数。

使用正则表达式实现年份的获取。选择“established”列，单击【编辑单元格】|【转换】选项，在打开的对话框中输入表达式“value.match（/.*（\d{4}）.*/）[0]”。其中，“.*”指的是一个零或多个字符序列（字母、数字、符号等），“\d”表示寻找的是一个数字，“{ 4 }”显示要恰好匹配 4 位数字。value.match( )函数的返回结果是数组，所以用“[0]”返回匹配的第一个元素，如图 3.120 所示。注意观察图的下方，“1947-four-year college”转换为“1947”，“1901 -”转换为“1901”，“Established

1985”转换为“1985”，然后数字化该列即可。

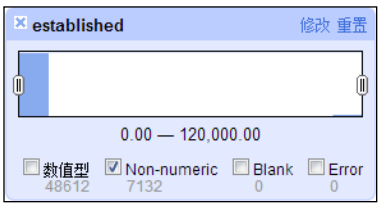


图 3.118　“established”列的数值归类

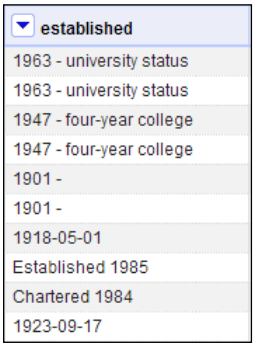


图 3.119　“established”列的非数值型数据

注意，归类中仍然存在 1352 条“established”列值为空的记录，可以根据需要决定保留还是删除这些空记录。

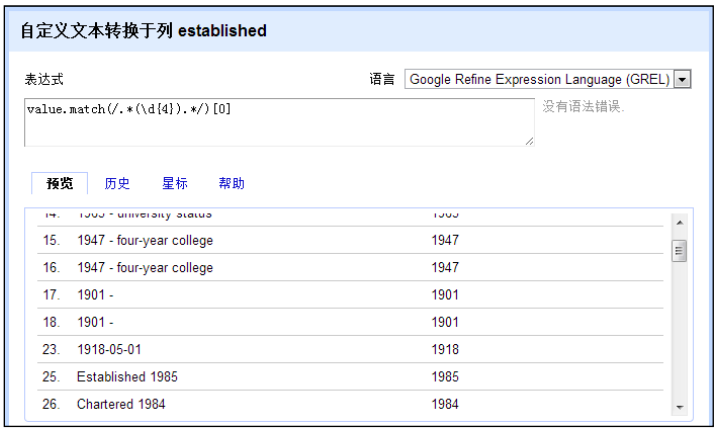


图 3.120　使用正则表达式获取“established”列的年份

（9）重复记录的问题。

单击“全部移除”按钮移除所有归类。仔细查看数据可以发现很多数据是重复的。为什么会出现这种情况呢？可能数据来源的时间不同，不同年份的学生的数量经常是不同的，也可能数据来源于不同的部门等。为保证数据的简洁，需要对重复行进行清理，为了使事情变得简单，可以仅保留每所大学排序第一的记录。

选择“university”列，单击“排序”菜单，排序依据是“文本”升序，“错误”和“空值”放在“合法值”后面，排序后的前 8 条记录如图 3.121 所示。

全部	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
24894.	Aarhus University	5270000000			Denmark	11000	1928	16395	17504	44
24895.	Aarhus University	5270000000			Denmark	11000	1928	16395	17504	32304
24896.	Aarhus University	5270000000			Denmark	11382	1928	16395	17504	44
24897.	Aarhus University	5270000000			Denmark	11382	1928	16395	17504	32304
48427.	Aarhus University	6196000000		NA	Denmark	11000	1928	16395	17504	44
48428.	Aarhus University	6196000000		NA	Denmark	11000	1928	16395	17504	32304
48429.	Aarhus University	6196000000		NA	Denmark	11382	1928	16395	17504	44
48430.	Aarhus University	6196000000		NA	Denmark	11382	1928	16395	17504	32304

图 3.121　排序后的前 8 条记录

选择“university”列，单击【编辑单元格】|【相同空白填充】选项，然后在排序后顶部出现的菜单中选择【Sort】|【固定行顺序】选项，如图 3.122 所示。

图 3.122　选择【固定行顺序】

操作后的记录如图 3.123 所示。对比图 3.121 和图 3.123 的记录，重复记录的“university”值用空白替代。

全部	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
1.	Aarhus University	5270000000			Denmark	11000	1928	16395	17504	44
2.		5270000000			Denmark	11000	1928	16395	17504	32304
3.		5270000000			Denmark	11382	1928	16395	17504	44
4.		5270000000			Denmark	11382	1928	16395	17504	32304
5.	Aarhus University	6196000000		NA	Denmark	11000	1928	16395	17504	44
6.		6196000000		NA	Denmark	11000	1928	16395	17504	32304
7.		6196000000		NA	Denmark	11382	1928	16395	17504	44
8.		6196000000		NA	Denmark	11382	1928	16395	17504	32304

图 3.123　相同空白填充后的前 8 条记录

选择“university”列，单击【归类】|【自定义归类】|【按空白归类】选项，选择“true”归类，再单击【全部】|【编辑行】|【移除所有匹配的行】选项。

单击“全部移除”按钮移除所有归类，查看清理后的记录。

思考：

上述方法删除重复记录是否符合你的要求，到底什么才是真正重复的记录？

注意：本步骤也可以用 Excel 的“删除重复项”功能实现。为实现以下步骤，需要先执行步骤 10 导出 Excel 数据后再操作。

打开导出的 Excel 数据，首先按“university”排序，查看记录重复的情况，然后在“数据”选项卡的“数据工具”中单击“删除重复项”，在打开的“删除重复项”对话框中单击“全选”按钮，即所有字段值完全一致才是重复记录，如图 3.124 所示，单击“确定”按钮完成数据清理。

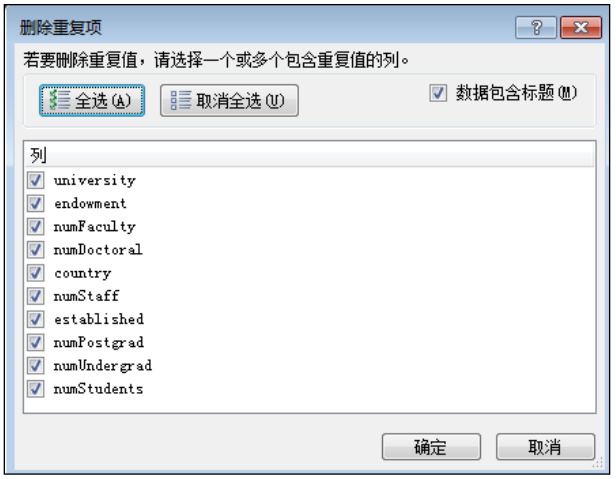


图 3.124　“删除重复项”对话框

（10）导出数据。

单击“全部移除”按钮移除所有归类。单击“导出”下拉按钮，选择“Excel”选项，在打开的对话框中设置导出的文件名称和位置后即可导出文件。具体操作参见 3.3.11 小节。

也可以导出包含所有操作的项目， 方便自己或他人学习。

思考：

尝试清理数据 data_Weibo.xlsx。

第 4 章

数据质量分析

- 数据合理性
- 游程检验
- 抽样分析
- 缺失数据的预测
- 时间序列预测

数据是数据新闻制作中最重要的基础条件，是数据探索和分析的前提。数据质量决定了数据新闻的质量，因此，数据质量的重要性无论如何强调都是不过分的，没有可信且高质量的数据，数据新闻无异于空中楼阁。

数据质量并没有一个统一的概念，但三个主要文献[1]均认为数据质量是数据适合使用的程度、是数据满足特定用户期望的程度。数据质量检查的目的就是保证数据的质量，简单地说，就是保证其正确性和有效性，为后续的数据分析、数据挖掘及可视化呈现等工作提供保证。如果从数据库的角度分析数据质量，评估维度主要包含以下六种[2]。

完整性（Completeness）。完整性用于度量哪些数据丢失了或者哪些数据是不可用的。

规范性（Conformity）。规范性用于度量哪些数据未按统一格式存储。

一致性（Consistency）。一致性用于度量哪些数据的值在信息含义上是冲突的。

准确性（Accuracy）。准确性用于度量哪些数据和信息是不正确的，或者数据是超期的。

唯一性（Uniqueness）。唯一性用于度量哪些数据是重复数据或者数据的哪些属性是重复的。

关联性（Integration）。关联性用于度量哪些关联的数据缺失或者未建立索引。

本章假设用户的数据已经手动或借助工具完成了数据清理，即已经解决了数据不完整、数据不一致、数据重复、数据存在错误、异常数据等问题，也可以认为从数据表面上看，不存在第 3 章涵盖的“脏数据”。

4.1 数据合理性

评估数据合理的方式主要有两种，一种是外部合理性检查，另一种是内部合理性检查。通过这两种检查尝试发现异常数据。

内部合理性。内部合理性是把数据与其自身进行比较，即与数据本身对照。这种检查可以通过直方图（Histogram）实现，也可以通过检查行数、确认数字加起来与其总和相符等方法判断。

外部合理性。外部合理性也称业务经验判断。外部合理性是把数据和其他数据进行比较，包括其他数据源、专家知识、以前版本的数据和依靠业务的相关知识及经验判断数据是否合理。

很多时候分析数据质量需要内、外部合理性综合分析，以全面判断数据的质量状况。

1　三个主要文献如下：

Dominik Lueebber,Udo Grimmer. systematic development of data mining based data quality tools[C].29 thvldb,2003；

Beverly K. Kahn, Diane M. Strong. Product and Service Performance Model for Information Quality: AnUp date. IQ 1998: 102-115,1998；

Cinzia Cappiello, Chiara Francalanci, Barbara Pernici.data quality assessment from user’spespective[C]. IQIS,2004。

2　http://baike.baidu.com/link?url=q9mRbCUgVkm5_2fR_hO-9JTsVpDdY13ICz3BK8n7LLtmi7HcLPIAUaO HnuxUqhfFMbYJ71NcOzxh5x2nnslJWq。

Excel 实现数据内外部合理性分析时，要使用“数据”选项卡上“分析”组中的“数据分析”工具，若没有该项，需要加载“分析工具库”宏程序，加载步骤如下。

（1）单击“Microsoft Office”按钮，然后单击“Excel 选项”。

（2）单击“加载宏”，然后在“管理”框中选择“Excel 加载宏”，单击“搜索”按钮。

（3）在“可用加载宏”框中选中“分析工具库”复选框，最后单击“确定”按钮。

4.1.1 内部合理性

直方图是内部合理性检查经常使用的方法，也是一种快速检查数据质量的重要技巧。

直方图也称质量分布图，是一种统计分析报告图，即由一系列高度不等的纵向柱形图或线段表示数据的分布状况。一般用横轴表示数据类型，用纵轴表示分布情况。直方图是统计分析方法[1]的核心。典型的正态分布是一种概率分布中央点最高，然后逐渐向两侧下降，曲线的形式是先向内弯，再向外弯的状态。检查数据的内部合理性就是判断数据直方图是否符合正态分布的特点。

制作直方图可以观察数据的分布频度，如果数据来源于产品，则可以判断生产过程是否稳定，预测生产质量。但需要注意的是，制作直方图的数据样本一般不少于 50 个，否则可能会产生较大的误差，可信度低，无统计意义。

制作直方图前，首先要对数据分组，分组涉及到组数和组距两个概念。组数是在统计数据时，把数据按照不同的范围分成几个组，组的个数称为组数。组距是每一组两个端点的差。组数和组距的选择决定直方图的质量。

案例 1：如已有某高中 2015 年高考成绩 5621 条（保存为 Excel 数据表），每条数据包含语文、数学和外语三个成绩。根据已有数据制作直方图，统计成绩的分布频度。

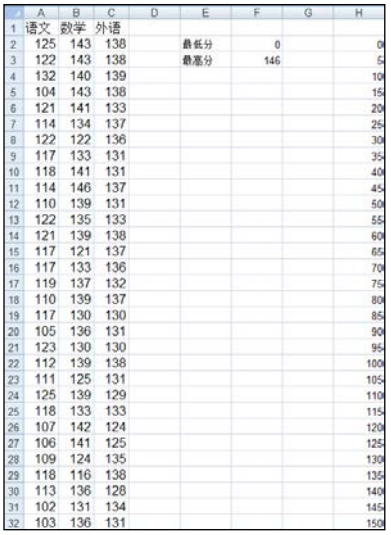
（1）设置区域名称。选择区域 A1:C5622（注意不选择首行列头），在名称框中输入名字“mark”，为区域 A1:C5622 添加名称，方便后期使用。

（2）确定组数和组距。使用函数 min（mark）计算最低分，用函数 max（mark） 计算最高分，分别为 0 和 146。高考单科满分是 150，而且考生较多，本案例组距设为 5，共 31 组，如图 4.1 所示。

（3）绘制直方图。在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“直方图”，单击“确定”按钮，如图 4.2 所示。

（4）设置直方图输入和输出选项。打开“直方图”对话框，“输入区域”是数据存储区，“接收区域”是组数区，因为“mark”不包含行头，所以不需要勾选“标志”，“输出选项”选择“新工作表组”，然后勾选“图表输出”复选框，如图 4.3 所示。

1 统计分析（statistical analysis）是指运用统计方法及与分析对象有关的知识，从定量与定性的结合上进行的研究活动。



	A	B	C	D	E	F	G	H
1	语文	数学	外语					
2	125	143	138		最低分	0		0
3	122	143	138		最高分	146		5
4	132	140	139					10
5	104	143	138					15
6	121	141	133					20
7	114	134	137					25
8	122	122	136					30
9	117	133	131					35
10	118	141	131					40
11	114	146	137					45
12	110	139	131					50
13	122	135	133					55
14	121	139	138					60
15	117	121	137					65
16	117	133	136					70
17	119	137	132					75
18	110	139	137					80
19	117	130	130					85
20	105	136	131					90
21	123	130	130					95
22	112	139	138					100
23	111	125	131					105
24	125	139	129					110
25	118	133	133					115
26	107	142	124					120
27	106	141	125					125
28	109	124	135					130
29	118	116	138					135
30	113	136	128					140
31	102	131	134					145
32	103	136	131					150

图 4.1　数据表、组数和组距

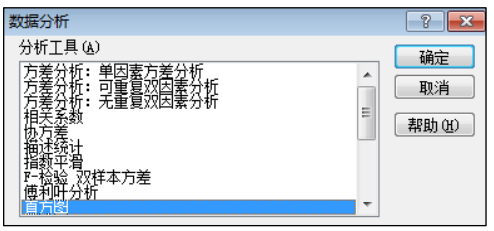


图 4.2　选择“直方图”分析工具

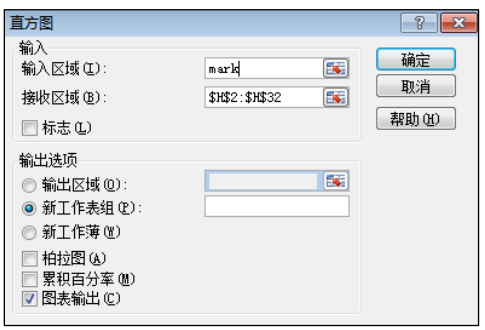


图 4.3　设置直方图

（5）查看显示的直方图。显示的直方图如图 4.4 所示。横轴显示接收组，纵轴显示频率情况。使用直方图不仅能够查看各组频数的分布情况，还可以查看各组之间的频率差异。本案例基本呈正态分布，数据的平均值决定了正态曲线的中心位置，方差决定了正态曲线的陡峭或扁平程度。数据内部检查合理。

组数和组距的选择对直方图的绘制影响较大，若组数和组距是{0，60，65，70，75，80，85，90，95，100，105，110，115，120，125，130，135， 140，145，150}，绘制的直方图如图 4.5 所示。对比图 4.4 和图 4.5 可以发现，图 4.4 正态分布明显，曲线呈钟型，两头低，中间高，左右基本对称；而图 4.5 由于组距的设置不是均匀的，{0，60}之间没有过渡，所以正态分布不明显，最左边的柱形因为包含的数据量大而呈现明显的凸起，若排除此柱形，查看其他柱形，正态分布还是比较明显的。

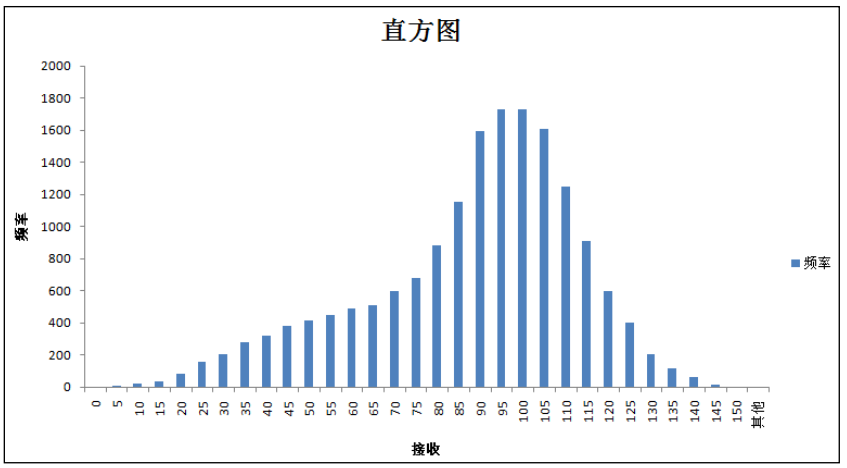


图 4.4　直方图效果 1

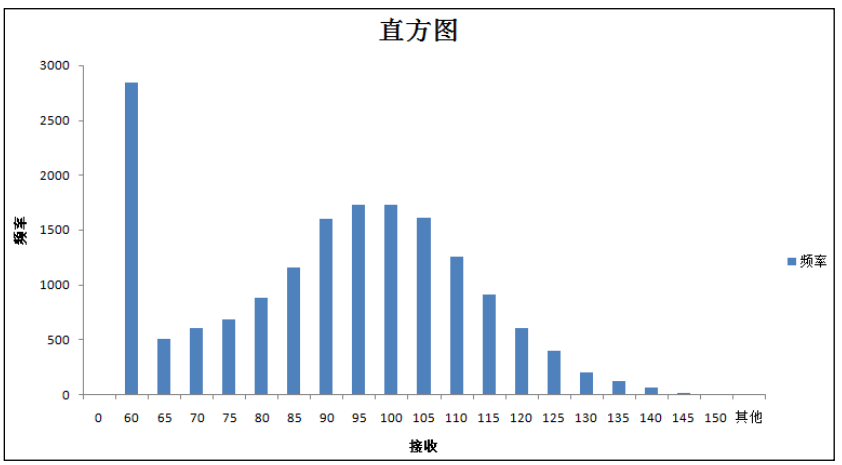


图 4.5　直方图效果 2

使用直方图也很容易发现有问题的数据。如图 4.6 所示，直方图中水平轴“105”处数据没有柱形，说明没有分数在范围（100,105）里，而根据经验（外部合理性中的业务判断）这是不可能的，特别是数据样本很大的时候。根据图 4.6 可以判定数据是不合理的，存在数据缺失的问题。针对不合理的数据制作数据新闻是没有任何意义的。

图 4.6 中最左侧水平轴“0”处呈现明显的凸起，说明分数为（0,5）的成绩过多，不符合正态分布，可能存在问题。注意，这与图 4.5 是不一样的，图 4.5 是组距不均匀导致的凸起，而图 4.6 的组距是均匀的。这也间接地证明了前面的分析，分值“105”的缺失，很可能是将分值为（100,105）范围的数据删除了或者修改为（0,5）。

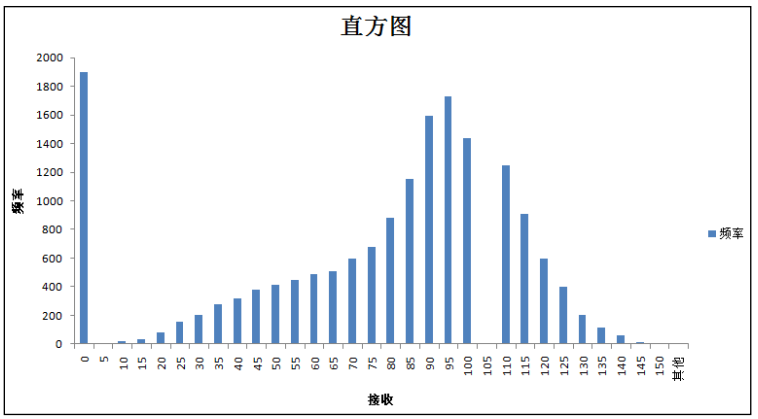


图 4.6　直方图效果 3

虽然通过直方图可以发现数据的不合理现象，并明确问题的缘由，但有些时候发现问题非常不容易，而且也无法确定问题是否一定存在，更无法判断何种原因导致的数据不合理。如图 4.7 所示，水平轴“105”处数据明显凹下，不符合正态分布。这可能是由于部分分数在范围（100,105）缺失，也可能是因为样本量较少而导致的。因为根据经验（外部合理性中的业务判断），范围是（100,105）的分数不应该显示得这么少；而且图 4.7 最左侧水平轴“0”处呈现明显的凸起也说明分值为（0,5）的成绩过多，不符合高考成绩常规（外部合理性中的与往年高考成绩的对比判断）。根据图 4.7 只能判定数据可能是不合理的，但无法确定是哪种问题导致的这种不合理。

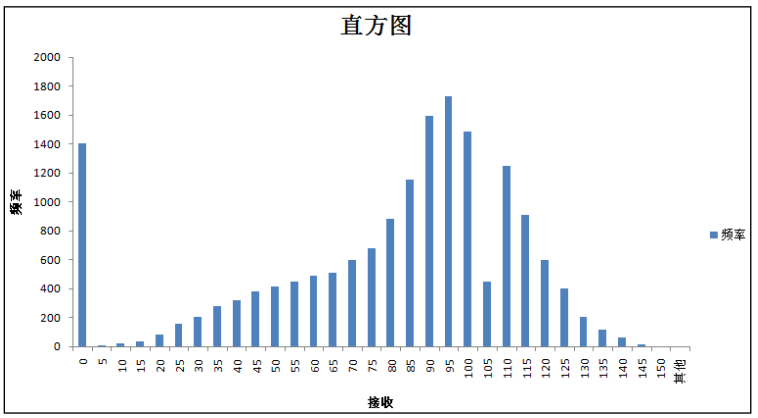


图 4.7　直方图效果 4

虽然使用直方图能发现存在问题的数据，但有问题的直方图不代表数据一定有问题。在英国《卫报》的一篇数据新闻“Animal testing: why the number of procedures is increasing”[1]中有一幅直方图（如

1　https://www.theguardian.com/news/datablog/2012/jul/10/animal-testing-risk-suffering。

图 4.8 所示），虽然该直方图并不符合正态分布，但该数据是真实的，而且数据样本量也不少，此数据来源于英国政府官方统计结果[1]。

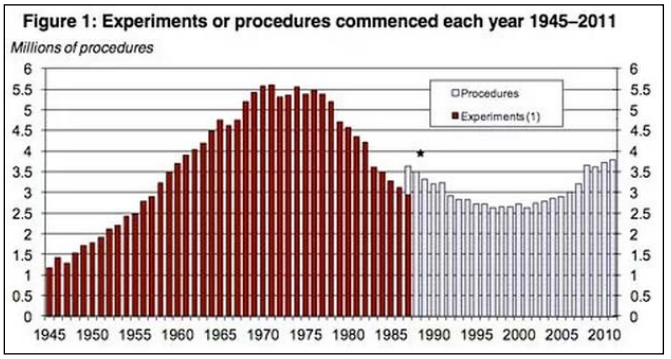


图 4.8 《卫报》的案例

还有一个《华尔街日报》关于俄罗斯选举的例子[2]，如图 4.9 所示。上面直方图的纵轴显示的是每个区投票给统一俄罗斯党（United Russia）的人数，下面直方图的纵轴显示的是每个区投票给其他党派（all other parties）的人数。从两个直方图上看，数据均可能存在问题，因为两个直方图最右侧的数据均偏高。尤其是上面的直方图，当投票率在 90%到 100%的范围时，投票给统一俄罗斯党的人数激增。因为我们无法像上个案例一样查找到数据来源，所以我们只能怀疑数据可能存在问题，但无法肯定数据一定是有问题的。

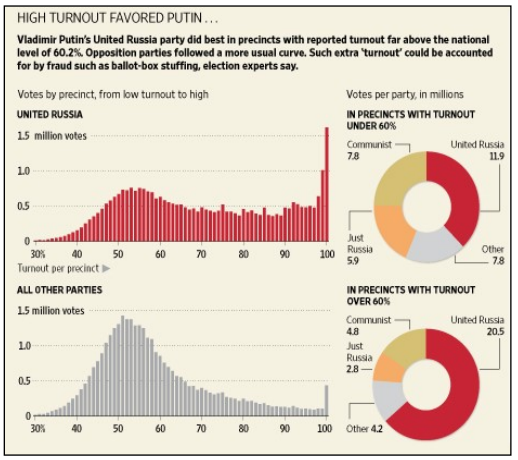


图 4.9 《华尔街日报》的案例

1 https://www.gov.uk/government/statistics/statistics-of-scientific-procedures-on-living-animals-great-britain-2011。

2 http://www.wsj.com/articles/SB10001424052970203391104577124540544822220。

4.1.2　外部合理性

为检查数据的外部合理性，经常使用 Excel 等工具为数据做描述性统计，然后根据描述统计结果与其他数据源、专家知识和经验判断数据是否合理。

Excel 的“描述统计”分析工具可以生成数据源区域中数据的单变量统计分析报表，提供有关数据趋中性和易变性的信息。此分析工具会涉及以下设置。

输入区域。定位准备分析数据区域的单元格范围。

分组方式。“逐列”表示输入区域中的数据是按列排列的，“逐行”表示输入区域中的数据是按行排列的。

标志位于第一行/列。如果输入区域的第一行中包含标志项（也称列头），则勾选“标志位于第一行”复选框；如果输入区域的第一列中包含标志项（也称行头），则勾选“标志位于第一列”复选框。如果输入区域没有标志项，则不勾选该复选框，Excel 自动在输出表中生成适宜的数据标志。

输出区域。定位输出结果左上角的单元格地址，用于设置输出结果的存放位置。

新工作表组。在当前工作簿中插入一个新工作表，并在新工作表的 A1 单元格开始存放统计结果。如果需要给新工作表命名，则在右侧文本框中输入名称。

新工作簿。创建一个新工作簿，并在新工作簿的第一个工作表中存放统计结果。

汇总统计。勾选此复选框，将计算并输出 16 个统计结果，包括平均、标准误差、中位数、众数、标准差、方差、峰度、偏度、区域、最小值、最大值、求和、观测数、最大（1）、最小（1）和置信度（95.0%）。

平均数置信度。若需要输出由样本均值推断总体均值的置信区间，则选中此复选框，然后在右侧的文本框中输入所要使用的置信度。

第 K 大/小值。如果需要在输出表的某一行中包含每个区域的数据的第 K 个最大值或最小值，则勾选此复选框，并在右侧的文本框中输入 K 的数值。

案例 2：如已有某高中 2015 年高考成绩 5621 条（保存为 Excel 数据表），每条数据包含语文、数学和外语三个成绩。使用 Excel 计算描述统计。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“描述统计”分析工具。

（2）在打开的“描述统计”对话框的“输入区域”文本框中输入“A1:C5622”（输入“A1:C5622”后系统自动显示为“$A$1:$C$5622”），选择“逐列”分组方式，勾选“标志位于第一行”复选框；“输出选项”选择“新工作表组”，勾选“汇总统计”、“平均数置信度”、“第 K 大值”和“第 K 小值”复选框。如图 4.10 所示，单击“确定”按钮。

（3）计算的统计结果如表 4.1 所示。其中，中位数是统计学的专有名词，代表样本中的一个特殊数值，该值可将样本集合划分为相等的两个部分。对于有限的样本集，首先将样本值按由高到低排序后找出正中间的样本作为中位数。如果样本集个数是偶数，通常取最中间的两个数值的平均数作为中位数。

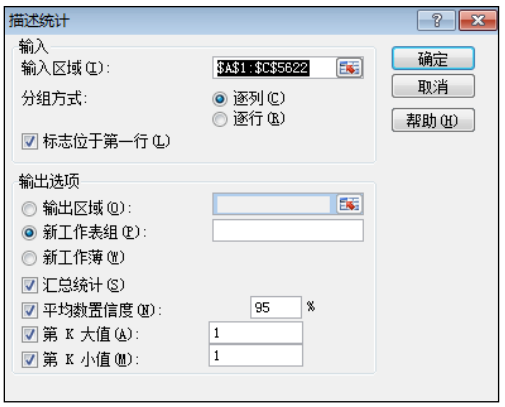


图 4.10 “描述统计”对话框

众数是在统计分布上具有明显集中趋势点的数值，代表数据的一般水平（众数可以不存在或多于一个）；也可以理解为众数是样本集中出现次数最多的数值，经常使用字母 M 表示。表 4.1 中“语文”成绩的众数是“99”，说明 5621 条成绩记录中，“语文”成绩出现频率最高的是“99”。

方差用来度量随机变量和其平均值之间的偏离程度。如果方差大，则样本集数据稳定性小，波动大；反之方差小，稳定性大。

标准差一般用来判定该组测量数据的可靠性，在正态分布中表现为正态分布曲线的陡峭程度。标准差越小，曲线越陡峭；反之，曲线越平坦。

偏度反映数据分布形态是否有对称性。若偏度是负数，数据位于均值左边的比位于右边的少，直观表现为左边的尾部相对于右边的尾部要长；因为有少数变量值很小，使曲线左侧尾部拖得很长。若偏度是正数，数据位于均值右边的比位于左边的少，直观表现为右边的尾部相对于左边的尾部要长；因为有少数变量值很大，使曲线右侧尾部拖得很长。若偏度接近于零，则认为分布基本是对称的，即两侧尾部长度对称。偏度用于检验数据是否呈正态分布。表 4.1 中“语文”成绩的偏度是“−0.60527”，说明曲线左侧尾部拖得比右侧尾部长。

峰度用于描述数据分布形态的陡缓，反映了尾部的厚度。正态分布的峰度是常数“3”，均匀分布的峰度是常数“1.8”。若峰度值小于 3，则分布具有不足的峰度。若峰度值大于 3，则分布具有过度的峰度。表 4.1 中“语文”成绩的峰度是“2.250681”，说明曲线峰度不足。

置信度 95%的值均大于 0.05，可以认为数据是真实而有效的。

表 4.1 描述统计结果 1

学科 / 项目	语　文	数　学	外　语
平均	95.3309	78.00427	84.87031
标准误差	0.148876	0.378039	0.369425
中位数	96	81	89
众数	99	87	111

续表

项目 \ 学科	语　文	数　学	外　语
标准差	11.16176	28.34287	27.69701
方差	124.5848	803.3181	767.1246
峰度	2.250681	–0.65849	–0.70812
偏度	–0.60527	–0.22483	–0.39682
区域	122	146	141
最小值	10	0	0
最大值	132	146	141
求和	535855	438462	477056
观测数	5621	5621	5621
最大（1）	132	146	141
最小（1）	10	0	0
置信度（95.0%）	0.291855	0.741103	0.724215

通过对比往年高考成绩的描述统计结果，可以发现数据是否存在问题。如对比表 4.1 和表 4.2（此高中某年高考成绩的描述统计结果），通过分析表 4.2 的统计结果查看数据是否存在问题。

首先查看平均值对比，表 4.2 的平均值比表 4.1 的平均值分别低 19.3、7.13 和 8.1，特别是语文成绩，根据业务经验可以判断数据不合理的可能性较大（也可能是当年的语文题目比往年难度大很多，导致学生整体分数下降）。中位数值的下降也进一步说明分数低的个数多，导致平均值的降低。

根据业务经验，一般高考的众数不会是“0”，即高考分数频度最高的不应该是“0”。而表 4.2 中“语文”、“数学”和“外语”的众数均为“0”，可以认为数据不合理的可能性较大（也可能是当年缺考学生特别多，或者作弊学生特别多等原因导致“0”分较多）。

表 4.2　描述统计结果 2

项目 \ 学科	语　文	数　学	外　语
平均	76.01797	70.87476	76.72923
标准误差	0.508889	0.449133	0.471064
中位数	90	75	83
众数	0	0	0
标准差	38.15308	33.67297	35.31721
方差	1455.658	1133.869	1247.305
峰度	0.134987	–0.43458	–0.36502
偏度	–1.31097	–0.38153	–0.59934
区域	132	146	141
最小值	0	0	0
最大值	132	146	141

续表

项目 \ 学科	语　文	数　学	外　语
求和	427297	398387	431295
观测数	5621	5621	5621
最大（1）	132	146	141
最小（1）	0	0	0
置信度（95.0%）	0.997618	0.880473	0.923467

4.2 游程检验

如果样本不是从总体中随机抽取的，那么所做的任何推断都将没有价值。很多时候数据新闻工作者需要判断获取的数据是否是随机的。例如，获取的调查问卷信息，如果是多个真实且随机的用户填写的，那么数据是随机序列。否则，如果一个用户填写了多份调查问卷，或者数据来自于其他的调查问卷，那么数据是伪随机的。游程检验是最简单且最常用的判断随机性的方法。

游程检验必须保证检验的变量类型是二分变量，如婚否、党员否等。游程检验的原则是：如果序列为真随机序列，则游程的总数应该不太多（如接近样本总数量）也不太少（如为 2）。否则就说明样本缺乏独立性，内部存在某种趋势或者有一定的联系；这可能是由于观察值间存在关联，或者样本来自不同的总体等原因。如数据转换的二分变量值是“00110111000100100010”。首先看“0”在这个序列中出现几次，假如有一个“0”，游程计数为一次游程，连续多个零也计数一次游程。本例中“0”为六次游程，“1”为五次游程，则数据“00110111000100100010”序列的游程是“11”。

案例 3：如已有某高中 2015 年高考成绩 5621 条（保存为 Excel 数据表），每条数据包含语文、数学和外语三个成绩，分别对三门成绩进行游程检验。

首先，制作二分变量。根据案例 2 可知语文、数学和外语三个成绩的中位数分别是 96、81 和 89。在 E2 单元格中输入公式“=IF（A2>=96,1,0）”，即语文成绩高于中位数 96 的为 1，否则为 0。将 E2 单元格的公式复制到 E3:E5622 区域，如图 4.11 所示。

查看“语文”成绩的二分结果，如图 4.12 所示。

	A	B	C	D	E
1	语文	数学	外语	语文的中位数	96
2	125	143	138		=IF(A2>=96,1,0)
3	122	143	138		=IF(A3>=96,1,0)
4	132	140	139		=IF(A4>=96,1,0)
5	104	143	138		=IF(A5>=96,1,0)
6	121	141	133		=IF(A6>=96,1,0)

图 4.11　制作“语文”成绩二分变量的公式

808	108	99	112		1
809	110	77	116		1
810	109	108	113		1
811	93	106	89		0
812	104	103	101		1
813	102	122	99		1
814	93	119	106		0
815	113	93	110		1
816	106	97	108		1
817	91	110	116		0
818	107	107	103		1
819	101	100	108		1
820	113	113	108		1

图 4.12　“语文”成绩二分变量结果

使用公式制作游程检验。在 G2 单元格中输入数字“1”，然后在 G3 单元格中输入公式“=IF（E3=E2,G2,G2+1）”，再将 G3 单元格的公式复制到 G4:G5622 区域，计算公式如图 4.13 所示。G 列最后一个单元格 G5622 的计算结果是“1856”，即“语文”成绩的游程是“1856”，该数值不太大也不太小，说明“语文”成绩是随机序列，成绩与顺序无关。

G
语文游程
1
=IF(E3=E2,G2,G2+1)
=IF(E4=E3,G3,G3+1)
=IF(E5=E4,G4,G4+1)
=IF(E6=E5,G5,G5+1)
=IF(E7=E6,G6,G6+1)
=IF(E8=E7,G7,G7+1)
=IF(E9=E8,G8,G8+1)
=IF(E10=E9,G9,G9+1)

图 4.13　计算“语文”成绩游程

4.3 抽样分析

大数据是当下流行的一种研究方法，意在不使用抽样调查方法，而是将所有的数据进行采集和分析处理。大数据具有 Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）和 Veracity（真实性）五个“V”的特点。

但在数据新闻制作中，很难真正地使用大数据，主要原因就是数据新闻的制作大多依赖于普通 PC，这就很难处理快速且数据类型多样的大量数据。而且当总样本数量过大时可能无法制作图表，或图表显示效果较差。所以，当我们获取大数据时，往往需要采用抽样分析的方法研究全部数据中的一部分样本。

抽样分析中最重要且最难的是如何保证抽取的样本能充分代表全部样本的属性和特征。Excel 提供的“抽样工具”可以通过“周期”或“随机”两种方法抽取样本。此分析工具主要涉及以下设置。

抽样方法。设置选取抽样的间隔。选择“周期”单选框并输入周期间隔，则按照间隔周期抽取样本数据；选择“随机”单选框将使用随机函数抽取样本数据。

样本数。在“样本数”文本框中输入随机值个数。每一个数值都是从输入区域的任意位置上抽出，并且任何数值都可以重复选取。

案例 4：如已有某高中 2015 年高考成绩 5621 条（保存为 Excel 数据表），每条数据包含语文、数学和外语三个成绩。使用 Excel“抽样工具”抽取 500 条语文成绩，分别计算描述统计并对比分析异同。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“抽样”分析工具。

（2）在打开的“抽样”对话框的“输入区域”文本框中输入“A2:A5622”，选择“随机”抽样方法，样本数输入“500”，“输出选项”选择“新工作表组”，如图 4.14 所示。

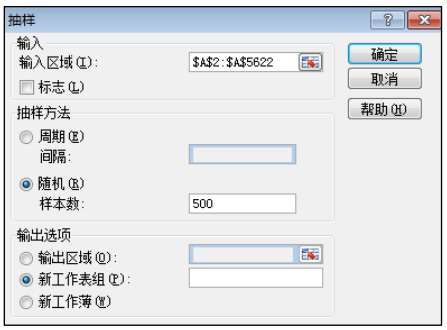


图 4.14 “抽样”对话框

使用 Excel 分别对两组数据做“描述统计”，如图 4.15 所示。可以看出，抽样数据与全数据差异不大，如“平均”、“中位数”、“众数”、“标准差”、“偏度”和“最大值”等。

	A	B	C	D	E
1	抽样语文			全部语文	
2					
3	平均	95.788		平均	95.3309
4	标准误差	0.487189		标准误差	0.148876
5	中位数	97		中位数	96
6	众数	96		众数	99
7	标准差	10.89387		标准差	11.16176
8	方差	118.6764		方差	124.5848
9	峰度	0.554818		峰度	2.250681
10	偏度	-0.52952		偏度	-0.60527
11	区域	74		区域	122
12	最小值	47		最小值	10
13	最大值	121		最大值	132
14	求和	47894		求和	535855
15	观测数	500		观测数	5621
16	最大(1)	121		最大(1)	132
17	最小(1)	47		最小(1)	10
18	置信度(95.	0.957194		置信度(95.	0.291855

图 4.15　全数据和抽样数据描述统计对比

案例 5：如已有某只股票的交易信息 3269 条，按周期抽样数据。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“抽样”分析工具。

（2）在打开的“抽样”对话框的“输入区域”文本框中输入“A2:I8”，选择“周期”抽样方法，设置间隔为“5”，设置“输出区域”为“K2”。周期设置为“5”是因为股票的交易是以 5 个工作日为一个周期。“抽样”对话框设置如图 4.16 所示。

（3）查看抽样结果，如图 4.17 所示。A 列是原始数据，K 列是抽取的数据。数据抽取时以 5 为周期，每隔 4 个数据抽取一个。

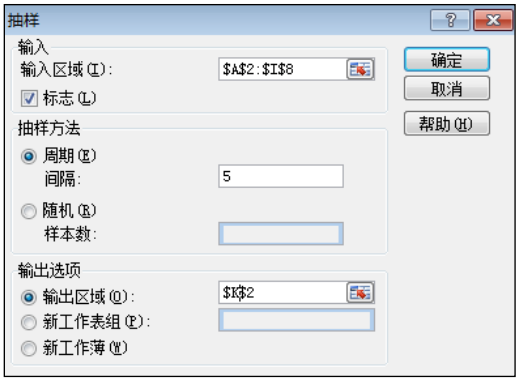


图 4.16　周期抽样设置

需要注意的是，本案例中输入区域是“A2:I8”，所以将对该区域的所有数据进行周期抽取，并最终显示为一列。在选择的输入区域中，每列包含 7 个数据，周期抽取按照从左到右的顺序抽取。如 A 列抽取第一个数据后，A 列剩下 2 个数据，为满足周期“5”的原则，抽取的第二个数据是 B 列的第三个数据。

如仅希望抽取所有开盘数据，则输入区域应该设置为“A2:A3270”。

	A	B	C	D	E	F	G	H	I	J	K
1	开盘	收盘	涨跌额	涨跌幅	最低	最高	成交量(手	成交金额(	换手率		周期抽取
2	250	261.21	8.96	0.0355	247.65	265.2	95707	246355	0.0084		253.96
3	239.8	252.25	16.94	0.072	234.13	255.85	81247	198907	0.0071		235.31
4	245.87	235.31	-8.66	-0.0355	234.02	248.49	67609	164600	0.0059		8.96
5	253.02	243.97	-11.12	-0.0436	240.6	254.8	61506	153084	0.0054		-6.94
6	253.96	255.09	1.14	0.0045	249	257.76	58389	147841	0.0051		-0.0436
7	259	253.95	-6.94	-0.0266	252.5	265.8	69304	178970	0.0061		234.13
8	266.98	260.89	-5.11	-0.0192	260.08	272	100014	266467	0.0088		260.08
9	262.12	266	5.94	0.0228	261	273	90910	243899	0.008		257.76
10	260	260.06	-0.99	-0.0038	257.77	264.5	52955	137841	0.0046		67609
11	255.64	261.05	2.02	0.0078	253.8	266.88	80462	209683	0.007		246355.38
12	265	259.03	-5.61	-0.0212	256.5	269.8	77852	204553	0.0068		178969.59
13	251.51	264.64	10.8	0.0425	249.8	267.54	96008	246274	0.0084		0.0054

图 4.17　周期抽样结果

4.4　缺失数据的预测

如果获取的样本数据存在数据缺失，则认为该数据是“脏数据”。但如果缺失的数据不多，且无其他数据文件可替代，可以尝试通过公开的数据网站、搜索引擎再次查询或再次众包二次收集数据；也可以使用工具根据一定的规则为样本添加缺失的数据。

“随机数发生器”分析工具可以选择某种分布产生的独立随机数来填充某个区域，可以通过概率分布来表示总体中的主体特征。此分析工具主要涉及以下设置。

变量个数。设置输出表中数值列的个数。

随机数个数。设置要查看的数据点个数，即生成符合条件的随机数的个数。每一个数据点出现在输出表的一行中。

分布。设置创建随机数的分布方法，包括“均匀”、“正态”、“柏努利”、“二项式”、“泊松”、“模式”和“离散”七种。其中，“均匀”分布以上限和下限来表征，其变量是通过对区域中的所有数值进行等概率抽取而得到的；“正态”分布以平均值和标准偏差来表征；“柏努利”分布以给定的试验中成功的概率（p 值）来表征，“柏努利”随机变量的值为 0 或 1；“二项式”分布以一系列试验中成功的概率（p 值）来表征，如可以按照试验次数生成一系列“柏努利”随机变量，这些变量之和为一个二项式随机变量；“泊松”分布以值 λ 来表征，λ 等于平均值的倒数，该分布经常用于表示单位时间内事件发生的次数；“模式”分布以下界、上界、步幅、数值的重复率和序列的重复率来表征；“离散”分布以数值及相应的概率区域来表征。

参数。设置用于表征选定分布的数值。

随机数基数。设置构造随机数的可选数值。

输出选项。设置输出位置，如“输出区域”、“新工作表组”和“新工作簿”。

案例 6：如已有某高中 2015 年高考成绩 5621 条（保存为 Excel 数据表），每条数据包含语文、数学和外语三个成绩。使用 Excel“随机数发生器”工具生成 16 条语文成绩。成绩符合正态分布，且平均值和标准偏差必须符合原始 5621 条数据。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“描述统计”分析工具取得全样本的平均值和标准差。

（2）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“随机数发生器”分析工具。在打开的“随机数发生器”对话框的“变量个数”文本框中输入“1”，“随机数个数”设置为“16”，“分布”为“正态”，“平均值”为“95”（来自于原始 5621 条数据的“描述统计”），“标准偏差”为“11.16176”（来自于原始 5621 条数据的“描述统计”），“随机数基数”为“210”，在“输出区域”文本框中输入“D2”，如图 4.18 所示，单击“确定”按钮。

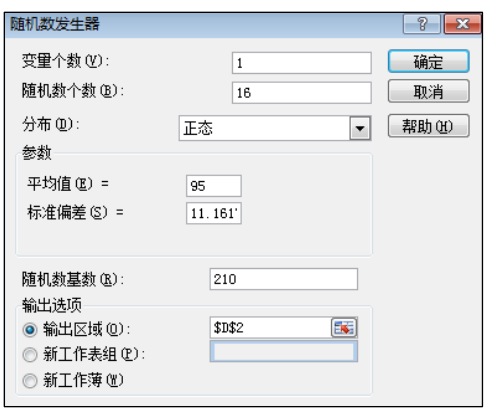


图 4.18 “随机数发生器”对话框

（3）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“描述统计”分析工具。计算原始 5621 条数据和 16 条随机数据的描述统计结果。

如图 4.19 所示，对比查看结果。图 4.19 左侧两列是全样本“描述统计”的结果，中间 1 列是获得的 16 个随机数，右侧两列是对 5621 条原始数据和 16 条随机数据共 5637 条数据“描述分析”的结果。二者差异很小，基本一致。

	A	B	C	D	E	F	G
1	语文			随机数		语文	
2				73			
3	平均	95.3309		82		平均	95.31308
4	标准误差	0.148876		83		标准误差	0.148685
5	中位数	96		89		中位数	96
6	众数	99		91		众数	99
7	标准差	11.16176		96		标准差	11.1633
8	方差	124.5848		86		方差	124.6192
9	峰度	2.250681		87		峰度	2.235779
10	偏度	-0.60527		83		偏度	-0.60165
11	区域	122		96		区域	122
12	最小值	10		87		最小值	10
13	最大值	132		83		最大值	132
14	求和	535855		110		求和	537279.8
15	观测数	5621		77		观测数	5637
16	最大(1)	132		107		最大(1)	132
17	最小(1)	10		98		最小(1)	10
18	置信度(95.	0.291855				置信度(95.	0.291481

图 4.19　对比查看结果

4.5 时间序列预测

很多社会现象是与时间相关的，研究随着时间的变化而变化的变量可以使用时间序列分析（Time series analysis）法。该方法基于随机过程理论和数理统计学方法，主要研究数据与时间序列的统计规律，是一种动态数据处理的统计方法。

Excel 中的“移动平均”、“指数平滑”和“回归”分析工具可以帮助用户解决这类问题。

4.5.1　移动平均

“移动平均”分析工具是基于特定的过去一段时间内变量的平均值，对未来值进行的一种数值预测。移动平均值提供了由所有历史数据简单的平均值所代表的趋势信息。此工具适用于变化较均匀的趋势预测。

Excel 提供的“移动平均”分析工具主要涉及以下设置。

输入区域。输入待分析数据区域的单元格范围。该区域必须单列存储，且至少包含四个数据单元格。

标志位于第一行。若输入区域的第一行中包含标志项（列头），则勾选此复选框。

间隔。设置移动平均计算时包含的数值个数。默认间隔为 3。

输出区域。设置输出区域左上角单元格的位置。如果勾选了“标准误差”复选框，则生成一个

两列的输出表，其中右侧列为标准误差值。如果没有足够的历史数据进行预测或计算标准误差值，则返回错误值#N/A。输出区域必须与数据源区域中使用的数据位于同一张工作表中。此对话框“输出选项”的“新工作表组”和“新工作簿”选项均为不可用状态。

图表输出。勾选此项，则在输出表中生成一个直方图。

标准误差。勾选此项，则输出表中包含标准误差的值。

案例 7：如已有一段时间（共 145 周）的某期货价格，预测期货价格走势。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“移动平均”分析工具。

（2）在打开的“移动平均”对话框的“输入区域”文本框中输入“B2:B146”，设置间隔为“9”，设置输出区域为“H2”，勾选“图表输出”复选框，如图 4.20 所示，单击“确定”按钮。

图 4.20 “移动平均”对话框

（3）查看结果。数据单元格 C146 的值显示为“3935”（若显示的数据带有小数，可以在“数字”选项卡中设置格式为整数），该值即为预测的趋势值。如图 4.21 所示是插入的直方图。

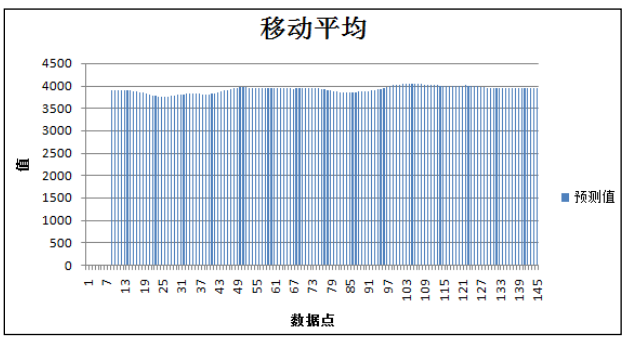


图 4.21 “移动平均”工具图表输出效果

间隔的设置对数值的趋势预测影响很大，在案例 7 中，“间隔”可以设置为 3 或以上的一个数值。可以使用“假设分析”工具求出最优时间间隔，具体步骤如下。

（1）在 C5 单元格中输入公式“=IF（A5<=F$2,"",AVERAGE（OFFSET（B5,-F$2,0,1））)”，并复制到 C6:C146 区域；在 F2 单元格中输入“3”，然后在 F3 单元格中输入“=SUMXMY2（OFFSET（B2,

F2,0,12-F2,1）,OFFSET（C2,F2,0,12-F2,1））/（12-F2）”。在 E4:E11 单元格中输入不同间隔的数值，利用数据表求得均方误差，结果如图 4.22 所示。

	A	B	C	D	E	F
1		某期货价格				
2	1	3911				3
3	2	3921				869
4	3	3877			1	452
5	4	3891	3911		2	659
6	5	3869	3921		3	869
7	6	3865	3877		4	1149
8	7	3879	3891		5	1265
9	8	3919	3869		6	689
10	9	3900	3865		7	439
11	10	3902	3879		8	343
12	11	3906	3919		9	249
13	12	3898	3900		10	277

图 4.22　使用“假设分析”工具求出最优时间间隔

（2）选中 E3:F11 单元格区域，单击“数据”选项卡上“数据工具”组中的“假设分析”，在打开的“数据表”对话框的“输入引用列的单元格”文本框中输入“F2”，如图 4.23 所示，单击“确定”按钮。F 列中求得最小值的均方误差是“249”，此时对应的间隔是“9”，即为最优时间间隔。

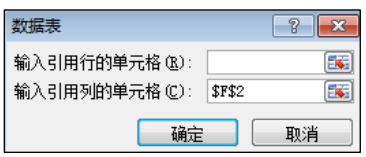


图 4.23　设置“假设分析”中的“数据表”

4.5.2　指数平滑

“指数平滑”分析工具基于前期预测值计算出新的预测值，并修正前期预测值的误差。工具中使用的平滑常数 α 决定了预测对前期预测误差的修正程度。

平滑常数 α 是[0,1]之间的小数。该值越接近于 1，远期实际值对本期平滑值影响程度的下降越迅速，预测变得越不稳定；该值越接近于零，远期实际值对本期平滑值影响程度的下降越缓慢。也可以理解为，数据波动越大，平滑常数 α 取值越大。数据波动平稳时，建议常数 α 取值小一些。实践中最常用的平滑常数是取值范围在[0.2,0.4]之间的小数。

确定平滑常数 α 取值的方法有两种，一种是经验法，另一种是计算法。

经验法依赖于时间序列的发展趋势和预测者的经验。当时间序列呈现较稳定的水平趋势时，平滑常数 α 的取值建议范围是[0.05,0.20]；当时间序列有波动，但整体趋势变化不大时，平滑常数 α 的取值建议范围是[0.1,0.4]；当时间序列波动很大，整体趋势变化幅度亦较大，呈现明显而迅速的上升或下降趋势时，平滑常数 α 的取值建议范围是[0.5,0.8]，以使预测模型灵敏度高些，能迅速跟上数据的变化；当时间序列数据是明显的上升或下降的发展趋势时，平滑常数 α 的取值建议范围是[0.6,1]。

计算法是根据数据具体情况，参照经验判断法确定平滑常数 α 的取值范围，然后通过分别计算 α 的不同取值，比较不同 α 值的预测标准误差，最终确定预测标准误差最小的平滑常数 α。

Excel 提供的“指数平滑”分析工具主要涉及以下设置。

输入区域。同“移动平均”分析工具的该选项，参见图 4.20。

阻尼系数。输入平滑常数 α 的取值。

标志。同“移动平均”分析工具的“标志位于第一行”复选项。

输出选项。同“移动平均”分析工具的该选项。

图表输出。同“移动平均”分析工具的该选项。

标准误差。同“移动平均”分析工具的该选项。

案例 8：如已有一段时间的某股票价格，预测股票的价格趋势。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“指数平滑”分析工具。

（2）在打开的“指数平滑”对话框的“输入区域”文本框中输入“B2:B3270”，输入阻尼系数“0.6”，设置“输出区域”为“I2”，勾选“图表输出”和“标准误差”复选框，如图 4.24 所示，单击“确定”按钮。

图 4.24 “指数平滑”对话框

图 4.24 中阻尼系数使用经验法确定为“0.6”，因为从数据上看，随着时间的变化，股票的价格明显呈现上升的发展趋势，结果如图 4.25 所示。

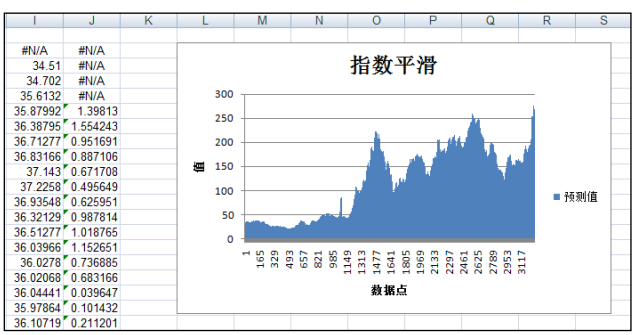


图 4.25 “指数平滑”工具计算结果

使用 Excel 的“规划求解”工具，通过计算可以求得最优阻尼系数，具体步骤如下。

（1）在 F2 单元格中输入平滑常数 α 的取值为“0.7”。

（2）在 C2 单元格中输入公式“=B2”，在 C3 单元格中输入公式“=$F$2*B2+（1-$F$2）*C2”。

（3）将 C3 单元格中的公式复制到 C4:C3270 区域，计算指数平滑值。

（4）在 F3 单元格中输入公式“=SUMXMY2（B2:B3270,C2:C3270）”，计算误差平方和，该值与标准误差同时达到最小。

（5）单击“数据”选项卡上“分析”组中的“规划求解”，在打开的“规划求解参数”对话框中设置目标单元格为“F3”，“等于”选项区的“最小值”为“0”，可变单元格为“F2”；单击“添加”按钮，在“约束”文本框中添加一个约束“F2<=1”，如图 4.26 所示。

图 4.26　“规划求解参数”对话框

（6）单击“选项”按钮，打开“规划求解选项”对话框，勾选“假定非负”复选框，如图 4.27 所示，再单击“确定”按钮。

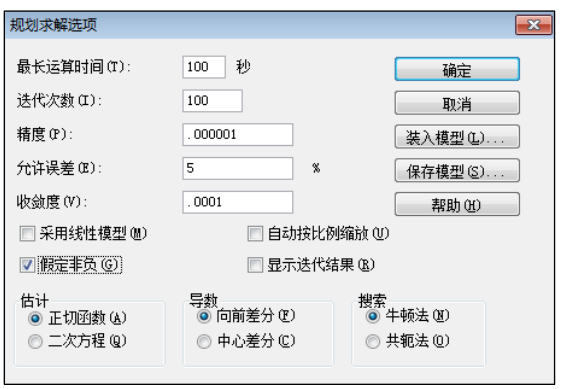


图 4.27　“规划求解选项”对话框

（7）返回“规划求解参数”对话框，单击“求解”按钮，计算得出最优平滑常数为 0.97。

4.5.3 回归

回归分析（regression analysis）是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

Excel 的“回归”分析工具可以完成时间序列分析，主要分析时间（自变量）和被分析变量（因变量）之间的线性关系。Excel 的“回归”分析工具主要涉及以下设置。

Y 值输入区域。设置被分析变量的数据区域。

X 值输入区域。设置自变量数据区域，此区域可以包含一个以上的变量，但变量必须是连续的区域。

标志。同“移动平均”分析工具的“标志位于第一行”复选项。

置信度。用于设置置信度，经常使用的置信度是“95%”。

输出选项。同“移动平均”分析工具的该选项。

残差图。在结果中嵌入一个正态概率图。

线性拟合图。在结果中嵌入一个线性拟合图。

正态概率图。在结果中嵌入一个正态概率图。

案例 9：如已有某国 54 年的 GDP 数据，预测该国 GDP 趋势。

（1）在【数据】菜单的【分析】选项卡中选择【数据分析】，在打开的“数据分析”对话框中选择“回归”分析工具。

（2）在打开的“回归”对话框的“Y 值输入区域”文本框中输入“C2:C55”，在“X 值输入区域”文本框中输入“A2:A55”，勾选“标志”和“置信度”复选框，将置信度设置为“95%”，输出区域为“F2”，如图 4.28 所示，单击“确定”按钮。

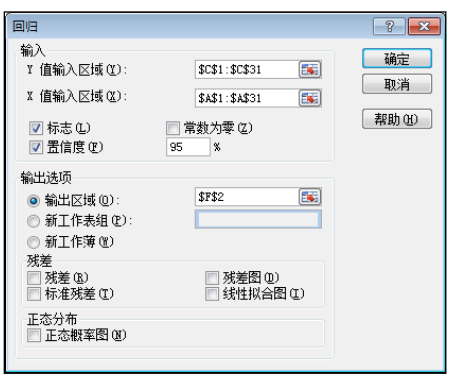


图 4.28 “回归”对话框

“回归”分析结果如图 4.29 所示。回归分析工具的输出结果包括回归统计表、方差分析表和回归参数表三个部分。

回归统计表包含的主要数据功能如下。

- Multiple R 表示相关系数，用来衡量自变量 *X* 与因变量 *Y* 之间相关程度的大小。本案例中

R=0.961508，表明二者之间的关系为高度正相关。

- R Square 表示复测定系数，也称拟合优度，是相关系数 R 的平方，用来说明自变量 *X* 解释因变量 *Y* 变差的程度，以测定因变量 *Y* 的拟合效果。本案例中 R=0.961508，则测定系数是 0.961508 的平方，即为 0.924497，说明用自变量 *X* 可解释因变量 *Y* 变差的 92.4497%。
- Adjusted R Square 表示调整后的复测定系数，本案例中该值是 0.9218，说明自变量 *X* 能解释因变量 *Y* 的 92.18%，因变量 *Y* 的 7.82%与其他因素相关。
- 标准误差表示拟合程度的大小，该值越小说明拟合程度越好。
- 观察值表示观察值个数。本案例中该值是 30 个。

SUMMARY OUTPUT

回归统计	
Multiple R	0.961508
R Square	0.924497
Adjusted R Square	0.9218
标准误差	349.7848
观测值	30

方差分析

	df	SS	MS	F	Significance F
回归分析	1	41946996	41946996	342.8458	3.03E-17
残差	28	3425784	122349.4		
总计	29	45372780			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-495.024	130.9849	-3.77924	0.000757	-763.334	-226.714	-763.334	-226.714
t	136.6157	7.378212	18.5161	3.03E-17	121.5021	151.7293	121.5021	151.7293

图 4.29　“回归”分析结果

方差分析表包含的主要数据功能如下。

- 第一列 df 表示自由度（degree of freedom）。其中，第一行是回归自由度，该值等于自变量数目，本案例中该值是 1；第二行为残差自由度，该值等于样本数目减去变量数量再减去 1，本案例中该值是 30−1−1=28；第三行为总自由度，该值等于样本数量减去 1，本案例中该值是 30−1=29。
- 第二列 SS 表示误差平方和，也称变差。其中，第一行是回归平方和，第二行是残差平方和，第三行是总平方和。
- 第三列 MS 表示均方差，是误差平方和除以相应的自由度得到的商。其中，第一行是回归均方差，第二行是剩余均方差。
- 第四列 F 值用于线性关系的判定。
- 第五列 Significance F 的 P-value 是 3.03E-17，远远小于显著性水平 0.05，故置信度达到 95%以上。表明该自变量对被解释变量有显著的影响，否则影响不够显著。

回归参数表主要用于回归议程的描述和回归参数的推断。表中第二行和第三行分别是截距和斜率的各项指标。根据回归参数表可以得出线性趋势方程如下：

$$Y = 495.024 + 136.6157 \times t$$

根据上述线性趋势方程，预测 2015 年和 2016 年该国 GDP 值分别是 3740.062 和 3876.678。

第 5 章

数据分析及可视化工具应用

- 数据可视化
- 数据可视化工具
- Tableau 下载和安装
- 创建第一个可视化作品
- 连接数据
- 数据视图
- 高级分析
- 仪表板
- 故事
- 作品发布
- Tableau 作品

人们在实践中发现，图像和图表是一种非常有效的传达信息与知识的方法。有研究表明，80%的人记得他们在自然界中用视觉所看到的信息，但只有 20%的人记得在书中阅读的文字内容 [1]。几个世纪以来，人们一直依赖于视觉表现，如早期的图表和地图，让人们更容易理解信息。

5.1 数据可视化

数据可视化（Data Visualization）是对数据的图像或图形格式的演示。通过数据可视化，数据新闻工作者能够看到数据背后的真相，尝试找到事件之间的相关性，并且通过简单易懂的方式呈现给读者。

数据可视化的重要性在于数据本身是难于理解的。想象一下，读者很难在一行接一行的数据中发现数据之间的规律和联系。数据可视化最大的好处是帮助读者更快地理解数据。数据可视化用另一种方式呈现数据，如在一个图表中突出显示某个大数据量，帮助读者快速发现关键点。尤其在现实的网络世界中，数据可视化可以更好地吸引网上冲浪者的眼球。

随着越来越多的数据的产生，数据可视化可以帮助用户分析信息，并提出一种让用户发现原本很难发现的模式和知识的方式。大量的数据是很难理解和接受的，数据可视化让这个过程变得更加容易。数据可视化适合展示大量的数据，如一张图表可能会突出显示多种不同的事项，读者可以在数据上形成不同的意见。

数据可视化提高了解释信息的能力。从海量的数据和信息中寻找联系并不容易，但是图形和图表可以在几秒内提供信息。

5.2 数据可视化工具

数据可视化软件是用于展示数据的工具，它选择正确的方式将数据用图像和图形展示，以达到最佳的可视化效果。

数据新闻制作对数据可视化工具提出了更高的要求。首先，我们希望工具具有实时性，即必须适应大数据时代数据量的爆炸式增长需求，数据图表制作迅速。其次，工具必须操作简单，因为大部分的新闻工作者没有太多的数学、统计分析和编程基础，只有功能丰富且易于操作的工具才能满足当下新闻人的需求。再次，支持多种数据格式和集成方式，因为数据的来源多种多样，只有支持团队协作数据、数据仓库或文本等多种方式才能适应数据新闻的需求。最后，输出便捷且能够通过互联网进行展现。

数据可视化软件用于传达并洞察复杂数据，本节我们尤其关注那些不需要任何编程基础就可以使用的软件。

1　来源于 http://www.cleverism.com/winning-data-visualisation-techniques/。

1．Tableau

Tableau 是一款数据可视化工具。无须编程就可以创建地图、条形图、散点图和其他图形。Tableau 将数据运算与美观的图表完美地嫁接在一起。

Tableau 界面清晰、易于操作，适合没有统计和编程基础的用户使用。该工具对海量数据处理非常快，图表创建迅速；图形种类繁多，嵌入了地图，还包含统计预测和趋势预测；仪表板和动态数据更新快捷，所见即所得；数据源丰富且输出方便，容易共享。

Tableau产品丰富，针对不同人群设置了多种版本，分别附带不同程度的支持和功能。如果首次使用该软件，笔者建议从Tableau Desktop入手，并申请Tableau public账号。Tableau工具可以在其官网[1]下载 15 天试用版，学习成本低。该工具对高校学生还有 1 年的免费试用期。Tableau作品如图 5.1 所示。

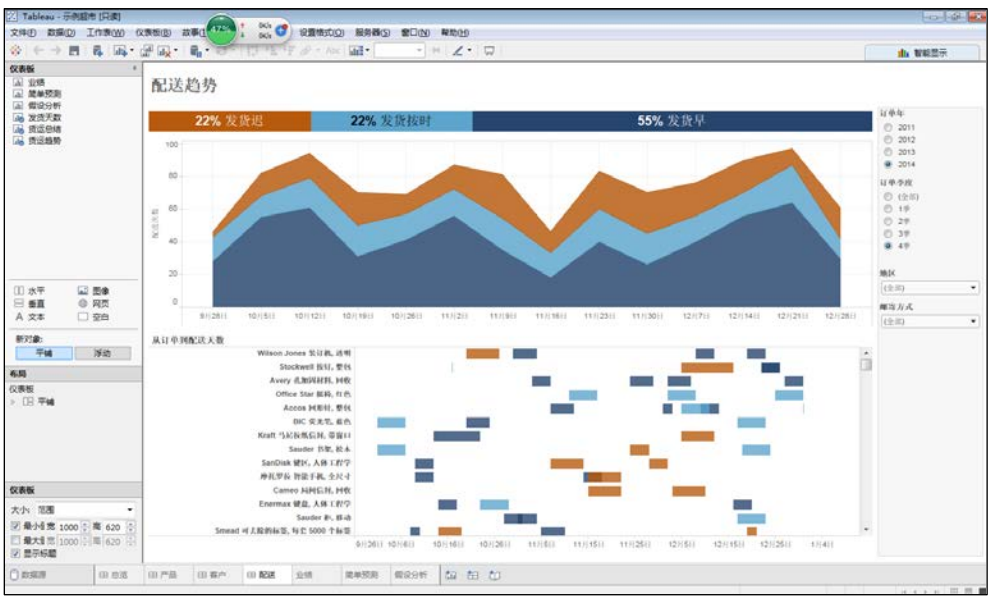


图 5.1　数据可视化工具 Tableau 作品

2．CartoDB

CartoDB[2]是一个专门制作地图的数据可视化工具，帮助用户将手机或网站里的位置数据进行可视化处理，无须任何编程，直接转换成直观的地图。它不仅能处理地理空间数据，还提供了数据分析、可视化和讲故事的功能。其网站包含丰富的文档[3]和教程[4]。CartoDB作品如图 5.2 所示。

1　Tableau 中文官网 http://www.tableau.com/zh-cn。

2　官网 http://cartodb.com/。

3　http://docs.cartodb.com/。

4　http://docs.cartodb.com/tutorials.html。

3．Visual.ly

Visual.ly[1]允许用户从Twitter、Facebook和Google Plus等社交网站采集数据，支持多种可视化模板。如分析某个账号为你的朋友建立一个描述其足迹的信息图，或者描述你最喜欢的图片和文章等。Visual.ly是制作信息图的利器，其作品如图 5.3 所示。

4．Datawrapper

Datawrapper[2]是一个在线工具，它可以帮助用户创建交互式数据可视化效果。它是一个由Journalism++ Cologne开发的开源工具，可以在几分钟内创建可嵌入的图表。因为它是开源的，所以任何人都可以贡献自己编写的代码，软件一直在不断的改进和提升。它有一个非常棒的图表库，用户可以将作品保存到图表库，方便用户互相查看作品。Datawrapper作品如图 5.4 所示。

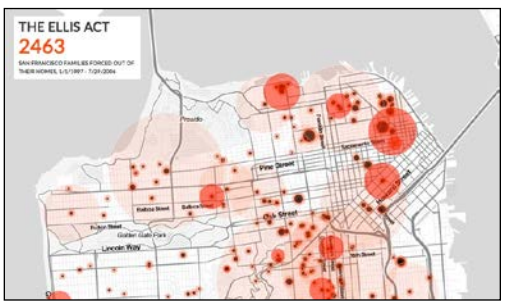


图 5.2　数据可视化工具 CartoDB 作品

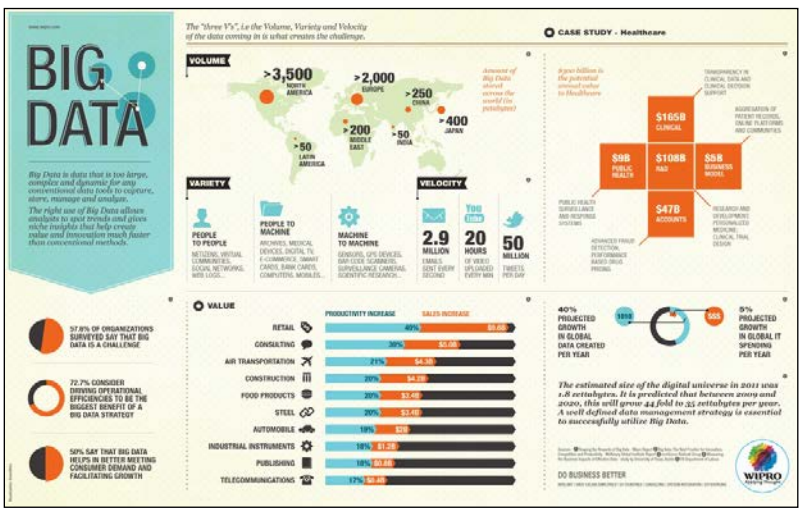


图 5.3　数据可视化工具 Visual.ly 作品

1　http://visual.ly。

2　https://www.datawrapper.de/。

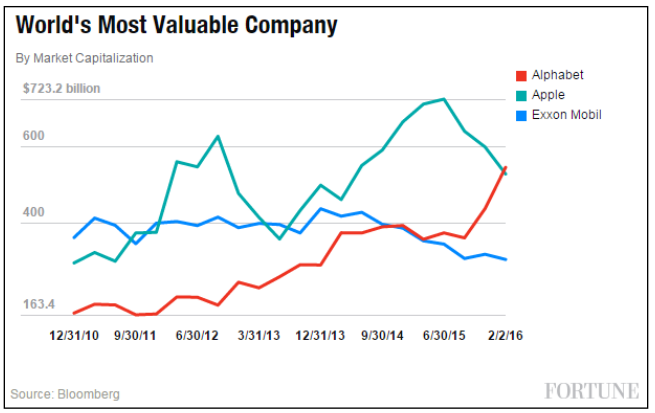


图 5.4　数据可视化工具 Datawrapper 作品

5.3 Tableau 下载和安装

Tableau 针对不同用户设计了多个版本，其版本信息如下。

- Tableau Server 是面向企业的版本，费用非常高。
- Tableau Online 是 Tableau Server 软件和服务的托管版本。
- Tableau Desktop 和 Tableau Public 是个人用户使用较多的两个版本。
- Tableau Reader 版本是一款完全免费的软件，用于打开和读取 Tableau 其他版本的数据和作品。

虽然版本不同，但基本功能类似。Tableau Public是完全免费的英文版本（目前没有中文版本），非常适合制作基于Web的交互式数据故事，可以方便地将作品发布在Tableau Public网站[1]上，方便任何人查看和下载。但Tableau Public的优点也是它最大的问题，因为所有的数据和可视化作品均上传到云服务器，对他人是公开的，所以完全公开的数据不适合某些需要数据保密的可视化工作；而且该服务器在美国，上传速度也是个问题。Tableau Desktop版本的数据和作品可以保存在本地计算机上，也可以发布到Tableau Public网站。

本节以 Tableau Desktop 版本为例，详细介绍 Tableau 的下载和安装步骤。

下载地址： http://www.tableau.com/zh-cn/products/desktop/download?os=windows。可以根据需要下载合适的 Mac 32-bit Windows 操作系统适用版本。下载的文件一般命名为“TableauDesktop-32bit-9-2-0.exe”，名称分为三个部分，首先是Tableau的Desktop版本，然后是 32 位Windows操作系统，最后是版本号 9.2.0。Tableau版本更新较快，读者下载的时候可能会有更新的版本。若下载的是Tableau Public[2]版本，则文件一般命名为 “TableauPublicDesktop-32bit-9-2-1.msi”，命名规则与Tableau Desktop版本类似。

1　https://public.tableau.com。

2　http://public.tableau.com/s/。

安装文件一般在 200MB 以下，双击安装文件安装即可，注意 Tableau Desktop 只有 15 天的免费试用期，试用期过后请付费购买并安装正版。Tableau Desktop Personal Edition 版本每人 999 美元，Tableau Desktop Professional Edition 版本每人 1999 美元，另外每年还需要缴纳一定的维护费用。

安装完成并连接到数据后，Tableau Desktop 的主工作区界面如图 5.5 所示。注意这是工作表的主工作区，仪表板工作区和故事工作区与此有很大区别，详细内容参见 5.8 节和 5.9 节。

图 5.5 中各区域的说明如下。

区域 1：菜单栏，包含 Tableau 的所有功能。

区域 2：工具栏，包含常用的功能，如撤销、重做和保存等，具体功能如表 5.1 所示。

区域 3：边条区，包含“数据”和“分析”两个选项卡，也称窗格。其中，“数据”窗格用于显示数据源、维度字段和度量字段等。“分析”窗格用于为图表添加分析信息，如汇总、模型和自定义等。

选择字段添加到行或列时，维度通常会产生标题。默认情况下，Tableau 将离散或分类的字段视为维度；将包含数字的字段视为度量，度量通常会产生轴。度量和维度也不是固定的，根据需要，有时候可以将度量转换为维度；如邮编信息默认是数值型数据，即度量，但该字段并不需要计算，所以可以将该字段拖动到维度中。

区域 4：标签栏，包含“数据源”、已经建好的“工作表”、“仪表板”和“故事”，以及“新建工作表”、“新建仪表板”和“新建故事”按钮。

区域 5：卡区，可以将数据拖放到该区域，并通过“页面”、“筛选器”和“标记”卡对图表进行设置。

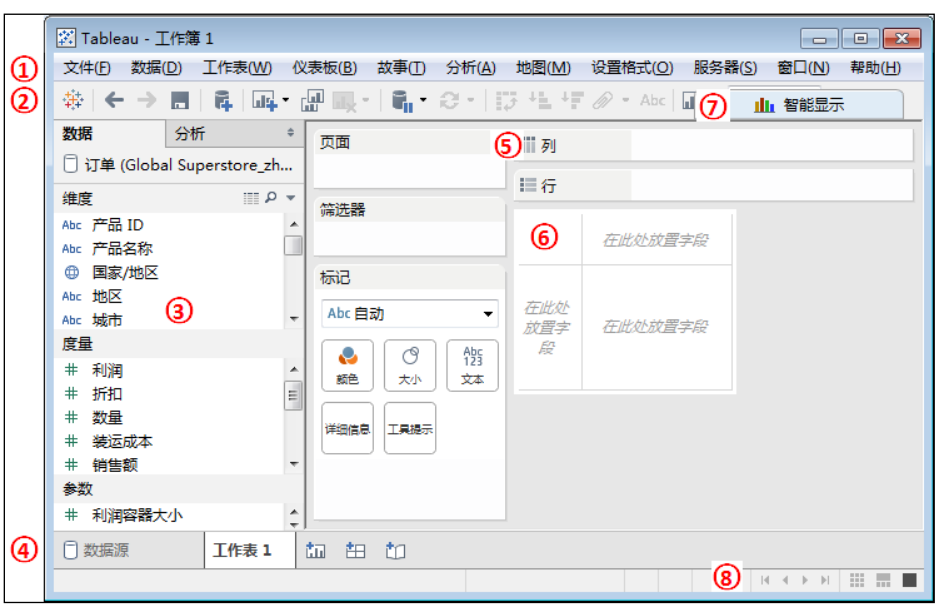


图 5.5　Tableau 主工作区界面

表 5.1　工具栏按钮及功能

按　钮	名　称	功　能
	Tableau 图标	转到开始页面
	撤销	撤销工作簿中的最近一次操作
	重做	重复使用“撤销”按钮撤销的最后一个操作
	保存	保存对工作簿所做的更改
	连接	打开“连接”窗格，可以在其中创建新连接，或者从存储库中打开已保存的连接
	新建工作表	新建空白工作表。使用下拉列表可创建新工作表、仪表板或故事
	复制工作表	创建与当前工作表完全相同的新工作表
	清除工作表	清除当前工作表。使用下拉列表可以清除视图的特定部分，如筛选器、格式设置、大小调整和轴范围
	自动更新	控制更改后的 Tableau 是否自动更新视图。使用下拉列表来自动更新整个工作表或只使用快速筛选器
	运行更新	进行手动数据查询，以便在关闭自动更新后用所做的更改对视图进行更新。使用下拉列表来更新整个工作表或只使用快速筛选器
	交换	交换“行”功能区和“列”功能区上的字段。每次单击此按钮，都会交换“隐藏空行”和“隐藏空列”设置
	升序排序	根据视图中的度量，以所选字段的升序来应用排序
	降序排序	根据视图中的度量，以所选字段的降序来应用排序
	组成员	通过组合所选值来创建组。选择多个维度时，使用下拉列表指定是对特定维度进行分组，还是对所有维度进行分组
	显示标记标签	在显示和隐藏当前工作表的标记标签之间切换
	显示/隐藏卡	显示和隐藏工作表中的特定卡。在下拉列表中可以选择要隐藏或显示的每个卡
	适合选择器	指定在应用程序窗口中调整视图大小的方式。可选择“标准适合”、“适合宽度”、“适合高度”或“整个视图”
	固定轴	在仅显示特定范围的锁定轴及基于视图中的最小值和最大值调整范围的动态轴之间切换

续表

按　钮	名　称	功　能
	突出显示	启用所选工作表的突出显示。使用下拉列表中的选项定义突出显示值的方式
	演示模式	在显示和隐藏视图（功能区、工具栏、“数据”窗格）之外的所有内容之间切换

每个工作表都包含可显示或隐藏的各种不同卡。卡是功能区、图例和其他控件的容器。如“标记”卡是用于设置标记属性的，它包含标记类型选择器及“颜色”、“大小”、“标签”、“详细信息”和“工具提示”等。

区域 6：画布区，也称可视化图表区。显示在设置区进行设置后的可视化图表。

区域 7：智能显示区。显示在设置区进行设置后可选择的图表类型。

区域 8：状态栏。显示当前视图下的基本信息和一些可选项。

5.4 创建第一个可视化作品

Tableau 是一个数据发现、数据分析和数据叙事的数据可视化平台，本节通过案例说明 Tableau 的基础功能，包含数据连接、创建多种图表、使用数据地图、工作表和仪表板的使用和输出等。

5.4.1 首次数据连接

Tableau 允许连接到多种格式的文本和数据文件，如常见的文本文件（.txt）、Excel 文件（.xls 或.xlsx）和 Access 文件（.mdb 或.accdb）等，详细内容参见 5.5.2 小节。连接一个 Excel 文件的步骤如下。

（1）打开 Tableau，在开始页面选择“连接到文件”中的“Excel”选项，如图 5.6 所示。在“打开”窗口中选择文件“Global Superstore_zh-cn.xlsx”后单击“打开”按钮。

（2）若连接数据时不在开始页面，可以单击【数据】|【新建数据源】选项创建数据连接，如图 5.7 所示；也可以单击工具栏中的第一个按钮“转到开始页面”（见表 5.1），或者直接按【Ctrl】+【2】快捷键。

图 5.6　选择“Excel”选项

图 5.7　选择【数据】|【新建数据源】选项

（3）在“连接数据页面”左侧的“工作表”区域中选择“订单”表并拖动到数据区，如图 5.8 所示。页面右下方将显示该表的字段和记录。

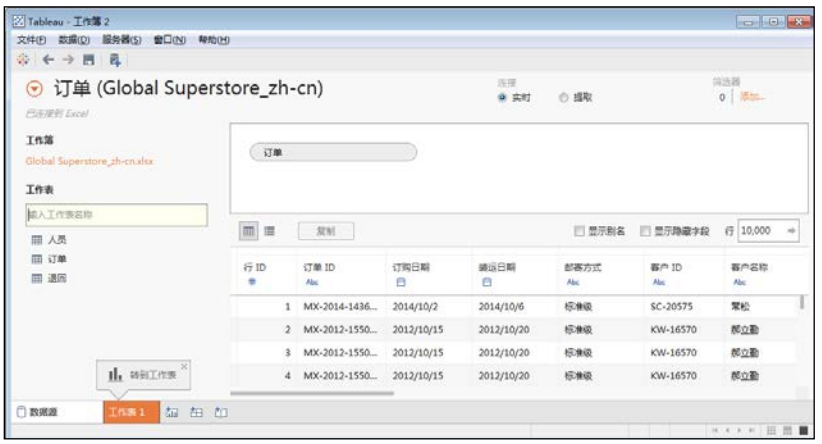


图 5.8　连接数据页面

注意，本例中的 Excel 文件“Global Superstore_zh-cn.xlsx”共包含“人员”、“订单”和“退回”三个表，本案例仅使用其中的“订单”表。

（4）单击图 5.8 所示的窗口左下角的“工作表 1”，进入 Tableau 主工作区。主工作区的数据包含“维度”和“度量”两部分。“维度”是离散的数据，如图 5.9 所示，如“国家/地区”。“度量”是连续的数据，如图 5.10 所示，如数量、利润和销售额。字段前面的符号表示数据类型，例如，“#”表示数值型数据，“Abc”表示文本型数据。

图 5.9 维度部分

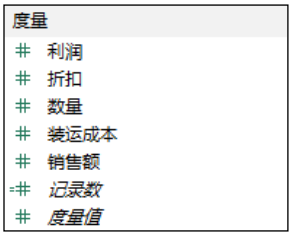


图 5.10 度量部分

5.4.2　首次创建多种图表

（1）创建一个数据地图，显示国家和销售额的关系。

“国家/地区”字段在“维度”区，默认为文本型数据，为制作数据地图，需要转换该字段的角色。用鼠标右键单击“国家/地区”维度字段，在弹出的快捷菜单中选择【地理角色】|【国家/地区】。注意查看转换角色后的“国家/地区”字段前面的符号，已经改变了。

单击“智能显示”使其打开。在“维度”区选择“国家/地区”字段，按住【Ctrl】键的同时在“度量”区选择“销售额”字段。然后在“智能显示”中单击“符号地图”。

（2）美化数据地图。在“维度”区选择“国家/地区”字段，拖动到“标记”卡中的“颜色”上，如图 5.11 所示，各个国家的圆形标记将呈现不同的颜色。

单击“标记”卡的“颜色”，设置透明度是 75%，并设置黑色边界，如图 5.12 所示。单击“标记”卡中的“大小”，增大标记到合适的大小。

图 5.11　“标记”卡

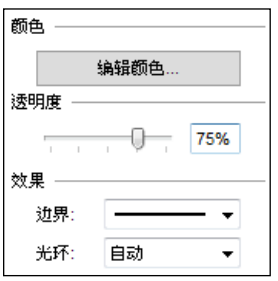


图 5.12　设置颜色

将“工作表 1”重命名为“各国家/地区销售总额”，最终效果如图 5.13 所示。

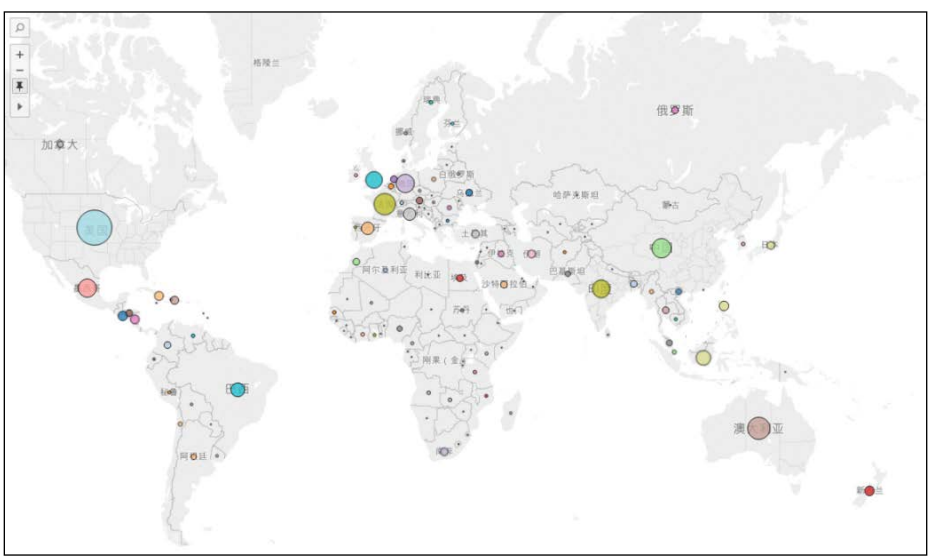


图 5.13　数据地图“各国家/地区销售总额”

（3）创建一个水平条图，显示类别和销售额的关系。

新建工作表“工作表 2”。将“销售额”度量字段拖到“列”功能区，再分别将“类别”维度字段、“子类别”维度字段拖到“行”功能区（注意顺序）。将“类别”维度字段拖动到“标记”卡中的“颜色”上。

在“维度”区选择“类别”字段，单击字段右侧的三角形按钮，在打开的菜单中单击【显示快速筛选器】。将“类别”和“子类别”设置为筛选器。将“工作表 2”重命名为“各类别销售额”，最终效果如图 5.14 所示。

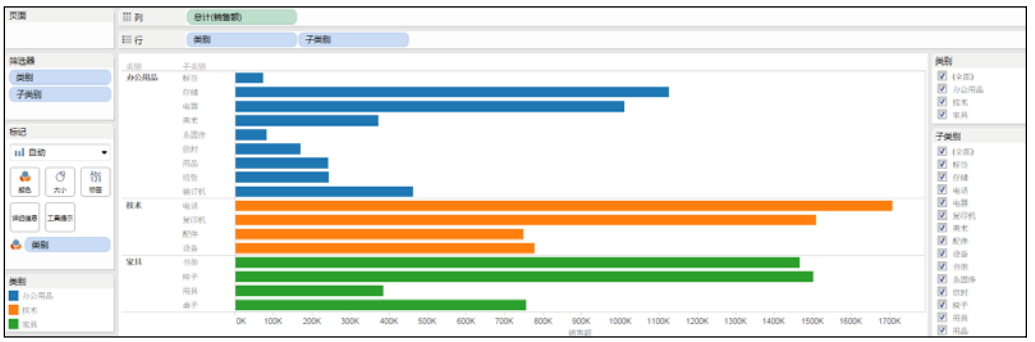


图 5.14　水平条图“各类别销售额”

注意，使用“筛选器”可以指定要包含或排除的数据。可以使用度量、维度或同时使用这两者来筛选数据。因为筛选器是独立的，所以在“筛选器”功能区上放置字段的顺序不会影响数据视图。如本例中“类别”和“子类别”筛选器的顺序不会改变结果。

（4）创建一个折线图，按月显示历年销售额的情况。

新建工作表“工作表 3”。将“订购日期”维度字段拖到“列”功能区，将“销售额”度量字段拖到“行”功能区（注意顺序）。单击“列”功能区上“订购日期”后面的三角形按钮，在打开的菜单中选择【月　　五月】，然后将“订购日期”字段拖动到“标记”卡中的“颜色”上。

将“工作表 3”重命名为“按月比较历年销售额”，最终效果如图 5.15 所示。创建图表的详细内容参见 5.5 节、5.6 节和 5.7 节。

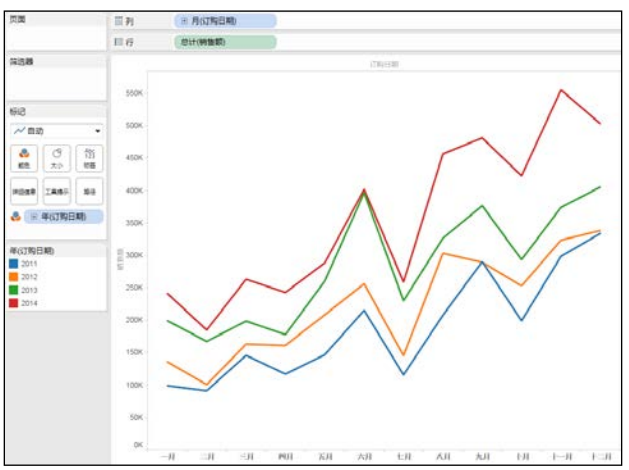


图 5.15　折线图“按月比较历年销售额”

5.4.3　首次创建仪表板

单击标签栏上的“新建仪表板”按钮，创建“仪表板 1”。将“各国家/地区销售总额”、“各类别销售额”和“按月比较历年销售额”三个工作表拖放到“仪表板 1”，删除“销售额”、“类别”和“国家/地区”图例。

编辑标题。选择“订购日期 年”图例，单击窗口上面的三角形按钮，在打开的下拉菜单中选择【编辑标题】，如图 5.16 所示，然后将标题修改为“年份”。

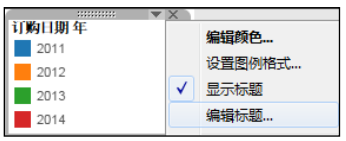


图 5.16　选择“编辑标题”

工作表联动。单击“类别”筛选器中的“办公用品”，仅有“各类别销售额”工作表显示了筛选后的数据，其他两个工作表没有实现筛选。为实现其他工作表的筛选，单击“类别”筛选器，再单击窗口上面的三角形按钮，在出现的下拉菜单中选择【应用于工作表】|【选定工作表】，勾选所有的工作表，如图 5.17 所示。仪表板最终效果如图 5.18 所示。

创建仪表板和故事的详细内容参见 5.8 节和 5.9 节。

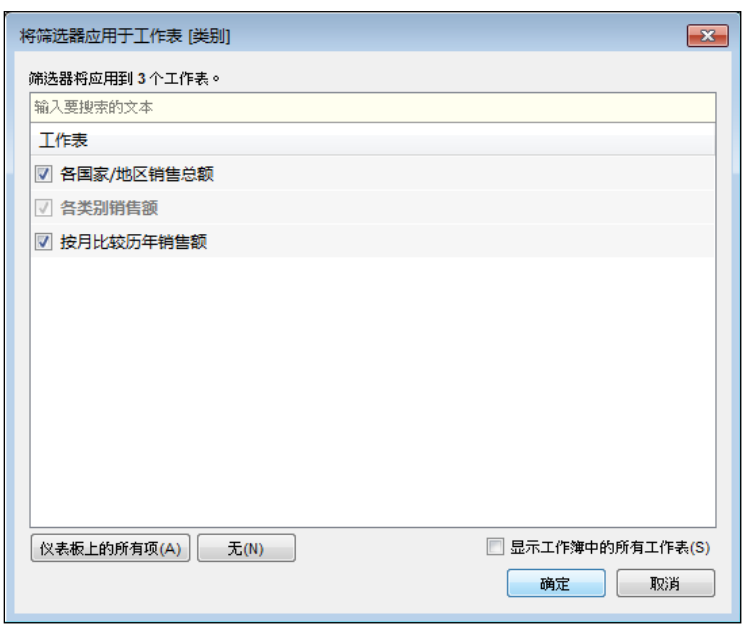


图 5.17　将筛选器应用于所有工作表

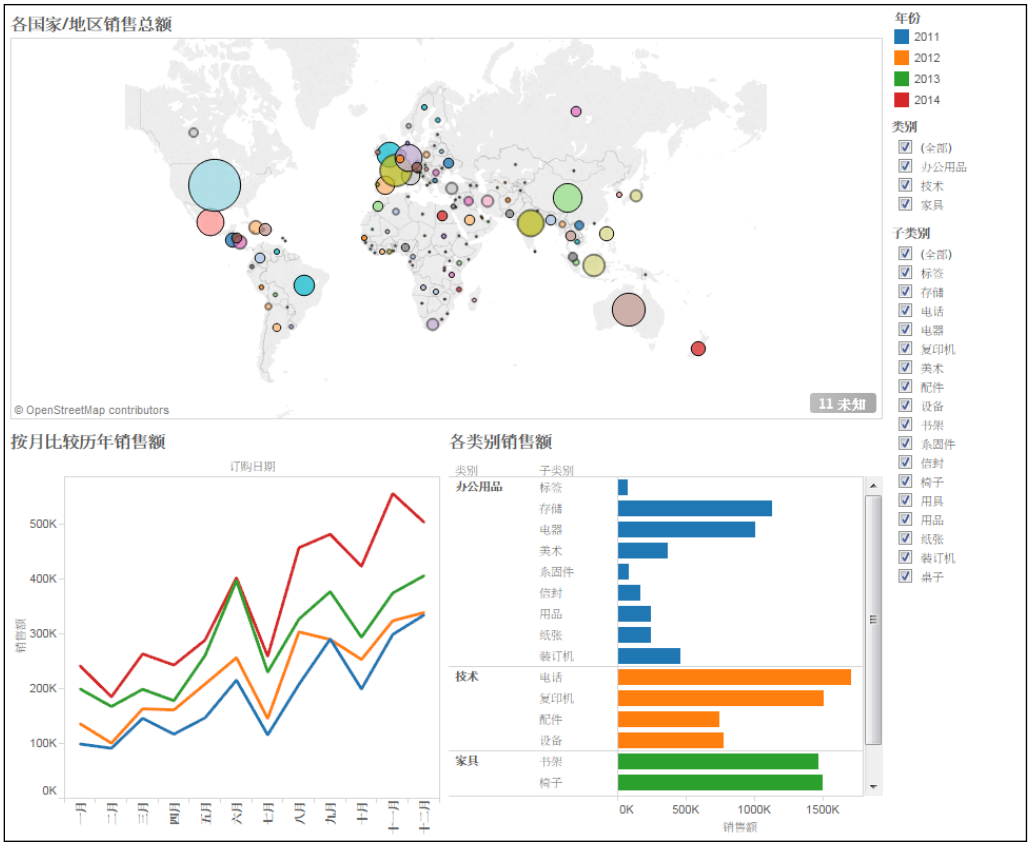


图 5.18　仪表板最终效果

5.4.4　首次输出

可以将 Tableau 作品输出为多种格式，最常见的格式是单击【文件】|【导出打包工作簿】选项后输出为 Tableau 特有的文件格式，该格式包含所有的数据、工作表和仪表板等，方便他人二次编辑。如将本案例打包工作簿保存为“5.4 创建第一个可视化作品.twbx”；也可以单击【文件】|【打印为 PDF】选项后输出为 PDF 文档，或者发布到 Tableau Public 网站，详细内容参见 5.10 节。

本章案例中还有一个遗留问题，即在仪表板的数据地图“各国家/地区销售总额”右下角显示“11 未知”，标明有 11 个国家或地区无法定位。单击“11 未知”，在出现的“编辑位置”对话框中编辑匹配位置，如“刚果共和国”匹配为“刚果（布）”、“刚果民主共和国”匹配为“刚果（金）”等，如图 5.19 所示，单击“确定”按钮即可匹配成功。

思考：

尝试新建三个工作表和一个仪表板呈现国家、地区、类别、订购日期与利润的关系，如图 5.20 所示。

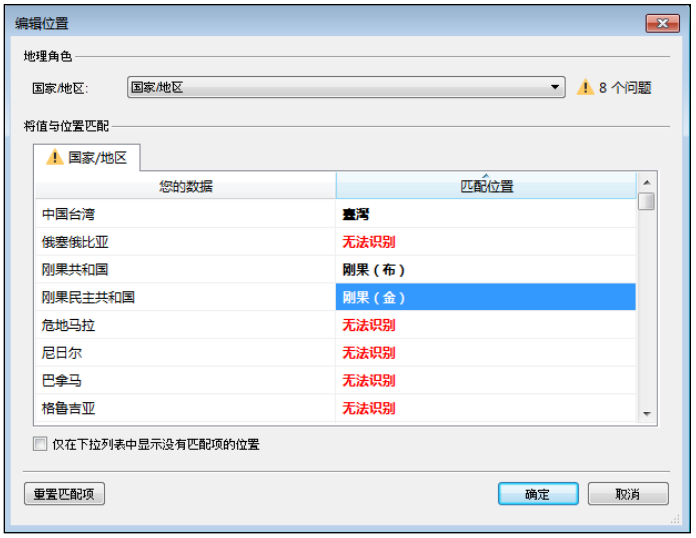


图 5.19　“编辑位置”对话框

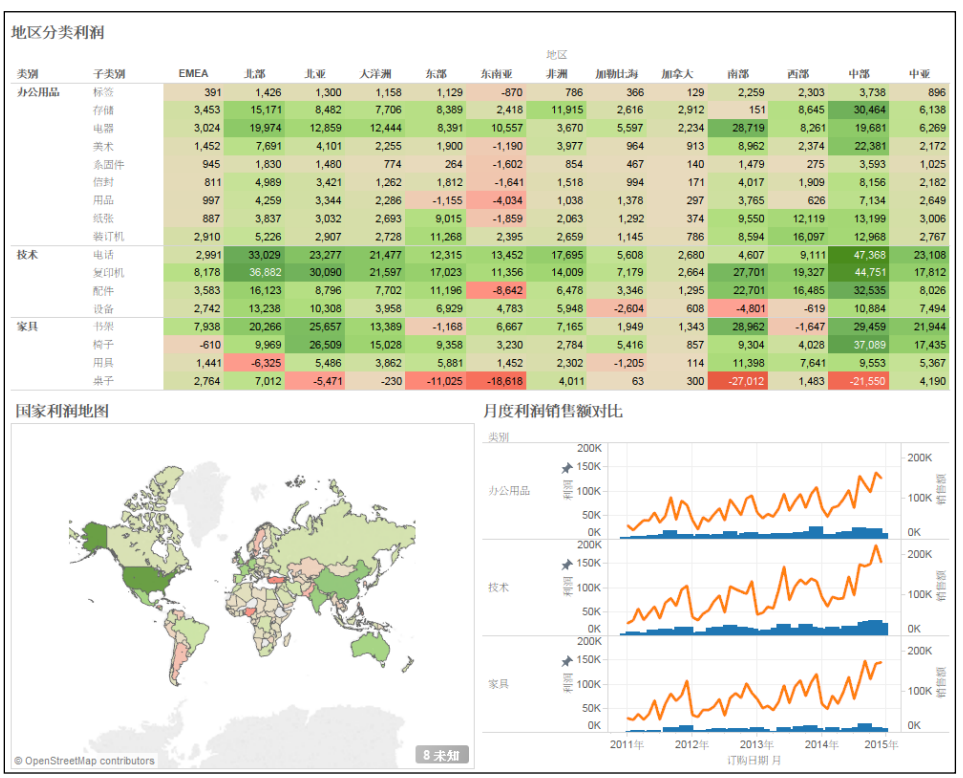


图 5.20　思考

第一个工作表是地区分类利润，按“类别”和“子类别”显示“地区”的利润总计，利润总计用红色和绿色颜色区分；绿色表示盈利，红色表示亏损。用“突出显示表”呈现效果。第二个工作表是国家利润地图，用“填充地图”呈现效果，按国家计算利润总和，并用颜色填充；颜色越绿表示盈利越多，颜色越红表示亏损越多。第三个工作表是月度利润销售额对比图，用折线图和条形图“双组合”按月呈现订购日期与类别、利润与销售额的对比。

可以尝试用自己喜欢的图表展现国家、地区、类别、订购日期与利润的关系，如加入筛选器或图例等。

5.5 连接数据

可视化效果依赖于数据，只有正确地连接数据才能保证可视化效果是正确的、美观的且有意义的。Tableau 可以连接一个或多个数据源，还可以提取数据。

5.5.1 在图表中查看数据

Tableau 用户不仅可以是专业分析师，还可以是非专业技术人员，如记者和编辑。使用 Tableau 可以轻松地实现数据的可视化、可交互实时展示与分析，这主要得益于 Tableau 的两个核心技术，一个是来自斯坦福的数据科学家独创的 VizQL 数据库，另一个是用户体验对易用性的完美呈现。二者结合使得 Tableau 在处理大规模、多维的数据时，也可以实时地从不同角度和设置下看到数据所呈现出的规律，图表让数据分析和数据挖掘变得平民化。

Tableau 使用 VizQL 数据库与其他各类源数据相连，并实现一定的操作。如图 5.21 所示是一个简单的水平条图，按类别显示利润。当把“类别”和“利润”分别拖动到行和列时，Tableau 自动产生一个 VizQL 查询，然后该查询转换为基于数据源的优化查询，如 SQL 查询。

普通用户无须关注 VizQL 查询、SQL 查询等，只需查看执行查询后的数据。选中图表中的“办公用品”，按【Ctrl】键选中其他两个类别，用鼠标右键单击某个列表条形图，在打开的快捷菜单中选择【查看数据】，在打开的对话框的“摘要”选项卡中可以查看数据摘要信息，如图 5.22 所示。

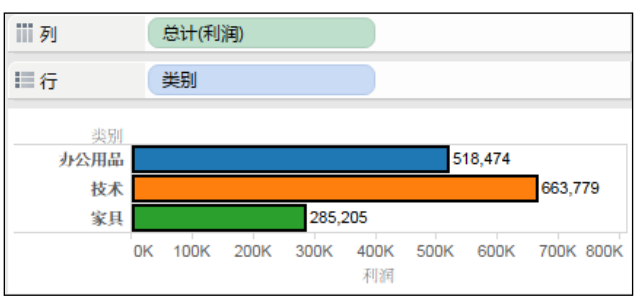


图 5.21 简单水平条图

图 5.22 “摘要”选项卡

在“基础”选项卡中可以查看数据的每条记录信息，若字段过多，可以取消“显示所有字段”复选框的勾选，仅显示行和列字段，如图 5.23 所示。

单击“全部导出”按钮可以将选中的数据导出为 CSV 格式的文件。

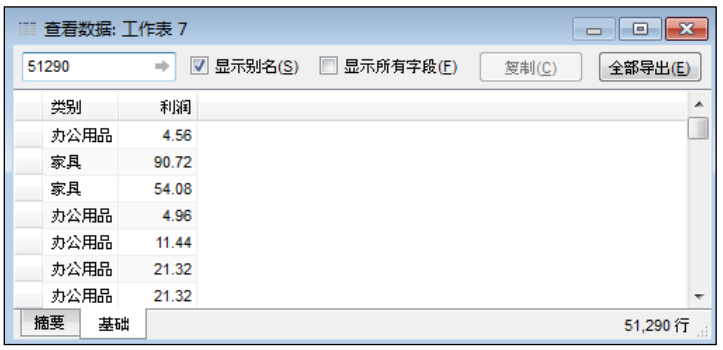


图 5.23　“基础”数据

5.5.2　简单数据连接

随着版本的升级，Tableau 对连接的数据源格式几乎没有限制，而且允许在一个工作表中使用多个不同格式的数据源。

建立数据连接的方式有以下三种。

- 在开始页面选择“连接到文件”。
- 单击【数据】|【新建数据源】选项，或者按【Ctrl】+【D】快捷键。
- 在标签栏中选择“数据源”，如图 5.5 所示的区域 4。

数据源包含三种，可以分别连接数据“到文件”、“到服务器”或“已保存数据源”。连接“到文件”可以连接不同种类的本地机或网络上的数据文件，包含以下五种格式。

- “Excel”包含.xls、.xlsx 和.xlsm 格式的文件。
- “文本文件”包含.txt、.csv 等格式的文件。
- “Access”包含.mdb 或.accde 等 Access 数据库文件。
- “统计文件”包含.sav、.rda、.rdata、.sas7bdat 等 SPSS 软件、R 语言、SAS 软件生成的数据文件。
- “其他文件”可以打开其他格式的文件，如 Tableau 工作簿文件（.twb）、打包工作簿文件（.twbx）、数据源（.tds）、打包数据源（.tdsx）、数据提取（.tde）等，还可以打开本地多维数据集文件（.cub）等。

如图 5.24 所示显示连接到文件“Global Superstore_zh-cn.xlsx”。

图 5.24　数据源界面

图 5.24 中各区域的说明如下。

区域 1：显示连接名称，双击可以修改。单击名称左侧的三角型按钮可以添加新的数据源。

区域 2：显示连接的数据类型和文件名。图 5.24 中连接的是 Excel 文件，显示“工作簿”；若连接到 Access 文件，则显示“数据库文件”。

区域 3：显示当前工作簿中包含的工作表。

区域 4：设置连接数据的方式，如“实时”或“提取”，连接本地数据多选择“实时”，“提取”的详细内容参见 5.5.5 小节。

区域 5：数据筛选器，单击“添加”按钮可以添加筛选器。

区域 6：连接区显示已用工作表的连接状态。连接到多个关系数据或基于文件的数据时，可以将一个或多个表拖到该区域并设置数据源。本例中选择了两个工作表，并且两个工作表自动进行了连接。

区域 7：左侧的“预览数据源”按钮用于显示字段及数据源中包含的前 10 000 行数据；“管理元数据”按钮用于以行方式显示数据源中的字段，方便用户快速检查数据源的结构并执行日常管理任务，如重命名字段或一次性隐藏多个字段等；还可以设置是否显示别名和是否显示隐藏字段等。

“到服务器”表示连接数据库中的数据或驻留在服务器上的服务。需要输入服务器名称和账号信息等登录到服务器，然后可以选择服务器的某个数据库的一个或多个数据表，如图 5.25 所示。

“已保存数据源”表示快速打开之前保存到“我的 Tableau 存储库”目录的.tds 格式的数据源。默认情况下，正常安装 Tableau 后系统已经提供了一些已保存数据源，如“示例—超市”。

在 Tableau 主工作区界面边条区的“数据”窗格中，用鼠标右键单击已经连接的数据，在打开的快捷菜单中选择【添加到已保存的数据源】选项，如图 5.26 所示，即将当前数据源保存，再次打开 Tableau 时会出现在“已保存数据源”中。

图 5.25　连接到服务器数据源

图 5.26　添加到已保存的数据源

5.5.3　连接多个数据源

Tableau 工作簿可以包含多个数据源的多个连接。每个数据源显示在“数据”窗格的顶部。每个工作表都有一个主数据源，可根据需要使用数据源来选择几个辅助数据源。主要数据源和辅助数据源通过指定的关系连接在一起。

如已有两个文件“Global Superstore_zh-cn.xlsx”和“利润计划.xlsx”。首先分别连接第一个文件的“订单”工作表和第二个文件的“利润预计”工作表，然后分别将两个文件添加到已保存的数据源，方法参见 5.5.2 小节的图 5.26。一个文件命名为“超市订单（Global Superstore_zh-cn）.tds”，另一个命名为“利润计划 （利润计划）.tds”。

运行 Tableau，在“已保存数据源”中打开“超市订单 （Global Superstore_zh-cn）”，然后单击工具栏上的“添加新的数据源”按钮，在“已保存数据源”中打开“利润计划 （利润计划）”，如图 5.27 所示。此时两个数据源无主次之分，并列显示在“数据”窗格中。若使用“超市订单 （Global Superstore_zh-cn）”数据源制作图表，则该数据源为主，而另一个数据源为次，注意“超市订单”数据源前面的对钩符号，如图 5.28 所示。

图 5.27　无主次之分的数据源

图 5.28　有主次之分的数据源

单击“超市订单 （Global Superstore_zh-cn)”，将该数据源中的“细分市场”拖动到“列”功能区，将“类别”和“总计（利润)”拖动到“行”功能区，再将“细分市场”拖动到“标记”卡中的“颜色”上，效果如图 5.29 所示。然后使用“利润计划 （利润计划)”数据源添加参考线，效果如图 5.30 所示。

图 5.29 单数据源效果

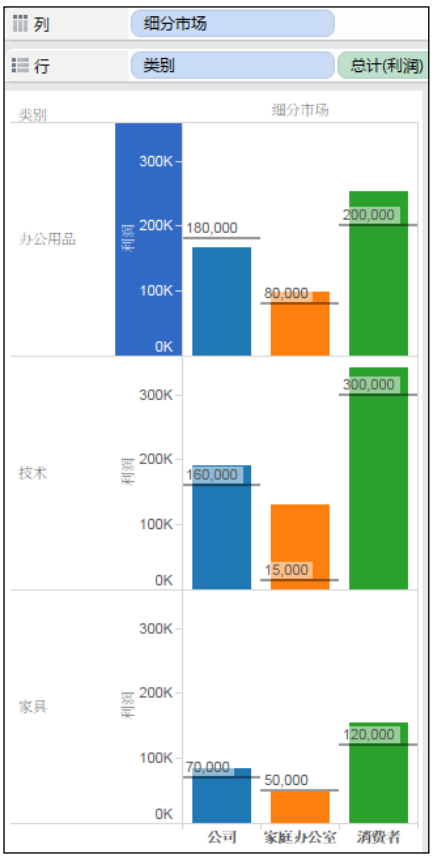


图 5.30 双数据源效果

单击“利润计划 （利润计划)”数据源，将“预计利润”拖动到“标记”卡中的“详细信息”上。用鼠标右键单击 *Y* 轴，在打开的快捷菜单中选择【添加参考线】选项，如图 5.31 所示。在打开的对话框中设置参考线的范围为“每单元格”，在“线”选项区设置“值”为“总计（预计利润)”，设置“标签”为“值”，如图 5.32 所示。

图 5.31　添加参考线

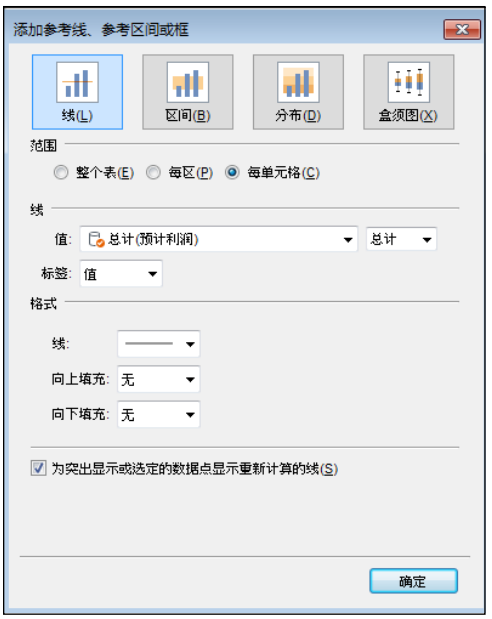


图 5.32　设置参考线

5.5.4　连接一个数据源的多个表

许多关系数据源都由与特定字段相关的表的集合组成。有时候我们需要使用数据源中的两个或多个表，例如文件“Global Superstore_zh-cn.xlsx”中包含三个数据表，分别是“人员”、“订单”和“退回”，为呈现每位员工的利润，需要使用“人员”和“订单”两个表，这两个表需要使用共有字段“地区”连接。

首先连接数据“Global Superstore_zh-cn.xlsx”，然后将“人员”和“订单”工作表拖入连接区。单击两个表的蓝色连接部分，打开“联接”对话框，连接的方式有“内部”、“左侧”、“右侧”、和“完全外部”四种，本案例选择“内部”连接，如图 5.33 所示。

四种连接方式的含义如下。

- “内部”连接是两个表的连接字段完全相同时才生成连接记录。
- “完全外部”连接显示左右两个表中的所有行。当某一个表中没有匹配的行时，则另一个表的选择列表包含空值（NULL），如果有则显示全部数据。
- “左侧”连接将左侧表（主数据源）中的所有数据与右侧表（次要数据源）中的数据进行匹配。当主数据源的特定成员不存在匹配项时，次要数据源的连接结果将为空值（NULL）。
- “右侧”连接将右侧表（次要数据源）中的所有数据与左侧表（主数据源）中的数据进行匹配。当次要数据源的特定成员不存在匹配项时，主数据源的连接结果将为空值（NULL）。

图 5.33　连接两个表

在主工作区“数据”窗格中查看两个表的字段。在维度中分别显示两个表名及其包含的字段，单击表名左侧的三角形按钮可以折叠或展开表字段，如图 5.34 所示。

图 5.34 “数据”窗格中的两个表

5.5.5　提取数据

提取数据是提高性能经常使用的一种方法，目的是为了保存数据源的一个子集。而且提取数据也提供了对数据的脱机访问，将数据提取到本地计算机方便用户使用。创建数据提取的步骤如下。

（1）打开“5.4 创建第一个可视化作品.twbx”，单击“数据”窗格上“维度”右侧的三角形按钮，在打开的下拉菜单中选择【隐藏所有未使用的字段】，如图 5.35 所示，则工作表中未使用的字段被隐藏了，这样可以限制记录的总数。

（2）用鼠标右键单击数据源，在打开的快捷菜单中选择【提取数据】，如图 5.36 所示。在打开的“提取数据”对话框中可以设置筛选器，还可以设置聚合以尽量减少数据提取量，如图 5.37 所示。单击“提取”按钮即可提取。提取数据仅包含可见字段，并且数据将按指定的方式聚合。

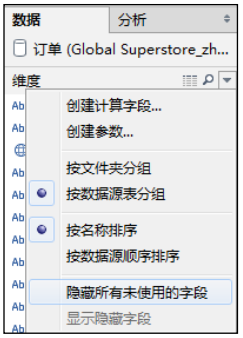


图 5.35　隐藏所有未使用的字段

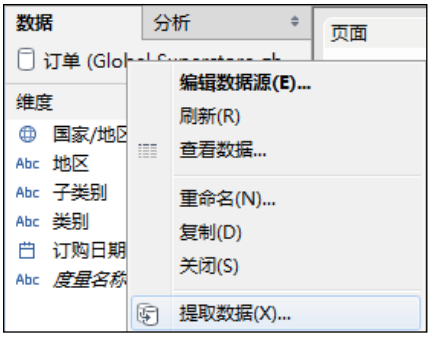


图 5.36　提取数据

注意数据提取后数据源图标的变化 。提取数据的时间与数据源的大小相关，若数据源很大，那么数据提取可能需要较长的时间，但提取数据并保存在本地计算机后，计算机性能将会大幅提高。

（3）用鼠标右键单击数据源，弹出快捷菜单，在【提取数据】子菜单中可以实现刷新、优化和追加数据等操作，如图 5.38 所示。

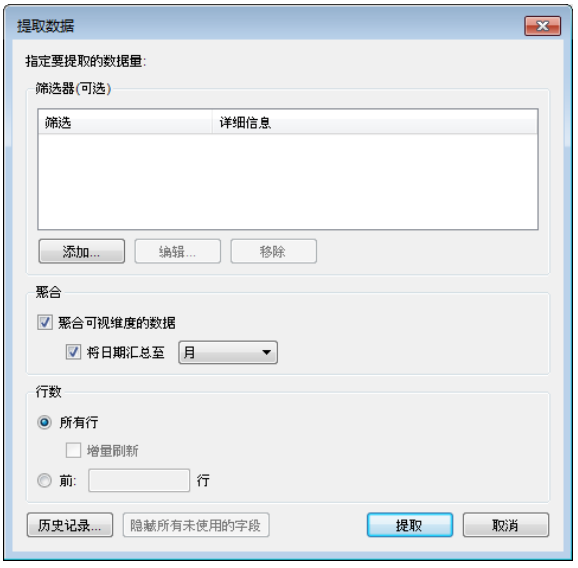


图 5.37　“提取数据”对话框

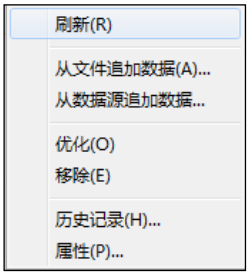


图 5.38　【提取数据】子菜单

（4）如需查看数据提取位置，可以在快捷菜单中单击【提取数据】|【属性】选项进行查看，如图 5.39 所示。

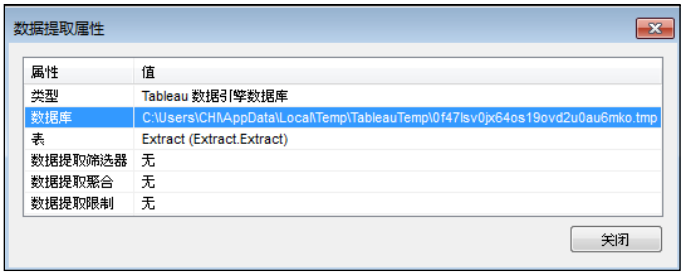


图 5.39　数据提取属性

5.5.6　数据类型

Tableau 支持文本、日期、日期和时间、数字、布尔和地理共六种数据类型。数据类型不同，图标也不同，详细内容如表 5.2 所示。

表 5.2　数据类型

图　标	说　明
Abc	文本值
	日期值
	日期和时间值
#	数字值
T|F	布尔值（仅限关系数据源）
	地理值（用于地图）

文本数据类型是以单引号或双引号开头和结尾的字符串文本，如“China”；日期数据类型以井号(＃)表示，如“#2020/1/15#”；日期和时间数据类型在日期后加上时间，如“#January 23, 1972 12:32:00 AM#”；数字数据类型表示数值，如 123、123.45 等；布尔数据类型也称为逻辑数据类型，只包含两种取值，即 true 或 false；地理数据类型是指如“北京”、“广州”或“东丰县”等地理位置。

5.6　数据视图

将现实生活中的数据转换成视觉线索，如颜色、形状、长度或位置等，目的是让读者可以通过一个简单的图表快速理解大量的数据。数据可视化时，要注意以下三个问题。

- **可视化的目的。**可视化能让读者理解什么？它回答了什么问题？

- **可视化的内容。**可视化包括哪些内容，即可视化中使用哪些字段?
- **可视化的结构。**究竟使用哪种图表适合，展示的是度量还是维度?

创建可视化数据视图是 Tableau 的重要功能之一，数据视图包含表组件和可选组件两部分。其中，表组件是视图的一部分，包括标题、轴、窗格、单元格和标记等。可选组件根据需要启用或禁用，如说明、字段标签和图例等。

5.6.1　工作表和工作簿

Tableau 使用了与 Excel 类似的工作表和工作簿文件结构。一个工作簿包含一个或多个不同类型的工作表，如工作表（也称视图）、仪表板或故事。工作表包含单个视图、功能区、图例和“数据”窗格。仪表板是多个工作表的集合，故事是多个仪表板的集合。

每次启动 Tableau 将自动创建一个空白工作簿，打开现有工作簿的方法有以下几种。

- 单击开始页面的工作簿缩略图（开始页面显示的是最近使用的工作簿）。
- 单击【文件】|【打开】选项可以打开.twb 或 .twbx 文件扩展名的工作簿。
- 双击一个工作簿文件。
- 将工作簿文件拖到 Tableau 图标上或拖到已打开的 Tableau 应用程序上。

Tableau 可以同时打开多个工作簿，每个工作簿显示在自己的应用程序窗口中。工作簿名称显示在 Tableau 的标题栏中。

创建工作表。工作表的创建可以单击【工作表】|【新建工作表】选项，也可以单击标签栏中的“新建工作表”标签，如图 5.40 所示的区域 4。

创建仪表板。仪表板的创建可以单击【仪表板】|【新建仪表板】选项，也可以单击标签栏中的“新建仪表板”标签，如图 5.40 所示的区域 5。

创建故事。故事的创建可以单击【故事】|【新建故事】选项，也可以单击标签栏中的“新建故事”标签，如图 5.40 所示的区域 6。

图 5.40　标签栏

区域 1：已创建的工作表。

区域 2：已创建的仪表板。

区域 3：已创建的故事。

区域 4：新建工作表标签。

区域 5：新建仪表板标签。

区域 6：新建故事标签。

隐藏工作表。工作表往往包含在仪表板或故事中，不单独使用。过多的工作表显示在标签栏中会导致查找不便且让标签栏显得复杂，可以使用隐藏功能隐藏工作表。单击工作表标签选择一个工

作表，或按【Ctrl】键选择多个工作表，用鼠标右键单击工作表标签，在打开的快捷菜单中选择【隐藏工作表】选项即可隐藏工作表。注意，只有在仪表板中使用的工作表才可以隐藏。

查看隐藏工作表。在仪表板【视图】菜单中选择【转到工作表】，或者单击仪表板中的“转到工作表”图标 均可以显示工作表。

取消隐藏工作表。若需要显示隐藏的工作表，可以用鼠标右键单击工作表标签，在打开的快捷菜单中选择【取消隐藏】选项即可取消隐藏工作表。

删除工作表。选择一个或多个工作表，用鼠标右键单击工作表标签，在打开的快捷菜单中选择【删除工作表】选项即可。注意，已经包含在仪表板或故事中的工作表是不能删除的，而且一个工作簿中始终至少包含一个工作表。

5.6.2 数据视图界面

数据视图界面如图 5.41 所示。其中，区域 1 是标题；区域 2 是字段标签；区域 3 是 *Y* 轴，也称行标题；区域 4 是说明；区域 5 是摘要；区域 6 是数据视图；区域 7 是 *X* 轴，也称列标题；区域 8 是图例。

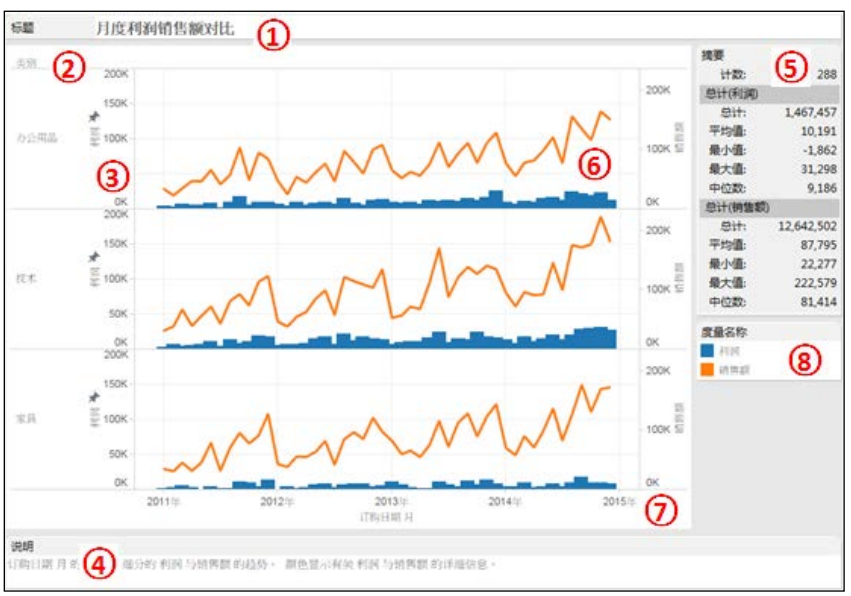


图 5.41 数据视图界面

编辑标题。用鼠标右键单击标题并在快捷菜单中选择【编辑标题】选项，或者双击标题本身，在打开的“编辑标题”对话框中输入新标题即可。使用对话框顶部的格式设置选项，可更改字体、颜色、样式和对齐方式。单击“插入”下拉按钮，选择下拉菜单中的选项可以添加页码、工作表名称和参数值等，如图 5.42 所示。说明和摘要的编辑方法与编辑标题类似。

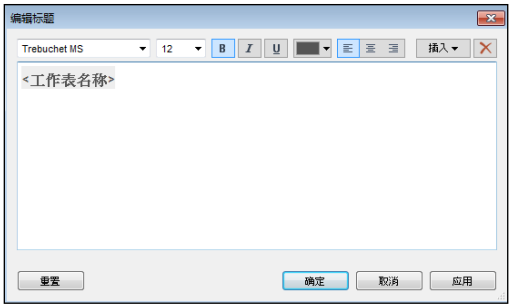


图 5.42　“编辑标题”对话框

编辑图例。为“标记”卡上的“颜色”、“大小”和“形状”添加字段时，会显示图例，以方便用户了解数据的编码方式。图例可以移动到合适的位置，也可以隐藏。在工作表中单击图例右上角的三角形按钮，在打开的下拉菜单中选择【隐藏卡】选项，也可以在仪表板中单击图例右上角的三角形按钮，在打开的下拉菜单中选择【从仪表板移除】选项，都可以将图例隐藏。

构建视图。可以手动构建视图，如拖动字段到“行”功能区和“列”功能区，也可以使用“智能显示”功能或“双击”自动生成视图。自动生成视图特别适合初学者，尤其在不确定使用哪种图形更适合的时候，而且自动生成视图节约时间。在制作复杂视图时一般建议先自动生成视图，然后手动深入细化。

5.6.3　文本表、压力图和突出显示表

文本表，也称交叉表或数据透视表。通过在“行”功能区和“列”功能区分别放置一个维度，然后将一个或多个度量拖到“标记”卡上的“文本”来完成视图的创建。

案例 1: 制作“地区”和“类别”文本表，如图 5.43 所示。

图 5.43　文本表

（1）连接数据。将数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）制作文本表。首先将“类别”拖动到“列”功能区，然后将“地区”拖动到“行”功能区，最后将“销售额”拖动到“标记”卡的“文本”上。

“列”功能区用于创建表列，“行”功能区用于创建表行。可以将任意数量的字段放置在这些功能区上。将维度置于“行”或“列”功能区上时，将为该维度的成员创建标题。将度量置于“行”或“列”功能区上时，将创建该度量的定量轴。向视图添加更多字段时，表中会包含更多标题和轴。“行”和“列”功能区上的内层字段决定默认标记类型。例如，内层字段为度量和维度，则默认标记类型为条形图。可以使用“标记”卡下拉菜单手动选择其他标记类型。

文本表的效果类似于 Excel，以数据为主，一般配合其他类型的图表放在仪表板中，主要目的是方便用户查看具体数据。

（3）突出显示表。为了能用颜色区分销售额的高低，可以使用突出显示表。单击“智能显示”，选择“突出显示表”查看效果，如果“列”和“行”功能区发生了转换，可以单击工具栏上的“交换”按钮（见表 5.1），效果如图 5.44 所示。

（4）转换为压力图。压力图用方块标明销售额的高低。本案例将“类别”拖动到“标记”卡中的“颜色”上进行区分，效果如图 5.45 所示。

	类别		
地区	办公用品	技术	家具
EMEA	276,686	300,855	228,621
北部	374,733	495,802	377,630
北亚	198,555	314,039	335,716
大洋洲	281,714	408,003	410,468
东部	205,516	264,974	208,291
东南亚	241,285	329,751	313,387
非洲	266,756	322,367	194,651
加勒比海	89,575	116,333	118,372
加拿大	30,034	26,299	10,595
南部	515,161	569,996	515,750
西部	220,853	251,992	252,613
中部	923,435	1,038,450	860,418
中亚	162,766	305,697	284,363

图 5.44　突出显示表

图 5.45　压力图

5.6.4　条形图

条形图是用户使用最广泛的图表类型之一，特别适合在类别之间比较数据，即用来比较不同分类的维度值（如性别、民族、城市、省份或部门等），但条形图往往不显示底层的数据细节。

制作条形图有两种方式，一种是直接制作条形图，另一种是将已经制作的其他图表转换为条形

图（当然已制作好的图表也可以转换为其他图表类型）。本小节使用Excel数据“省会城市空气质量.xlsx”，制作条形图来呈现一个月中各省会城市的天气情况。该数据来源于中华人民共和国环境保护部官方网站[1]的数据中心，案例获取了 2015 年 11 月 1 日至 2015 年 11 月 30 日全国 367 座重点城市每日空气质量报告（包含AQI指数、空气质量级别及首要污染物），然后筛选出省会城市。

案例 2: 制作简单条形图“AQI 均值排序”。

该条形图按 AQI 均值排序，方便用户查看空气质量均值最差和最优的省会城市。

（1）连接数据。数据源连接到文件“省会城市空气质量.xlsx”。

（2）制作基础图表。将“省会城市”维度字段拖动到“列”功能区，将“AQI 指数”度量字段拖动到“行”功能区。“行”功能区默认显示的是“总计（AQI 指数）”，此时呈现简单的条形图效果。单击“总计（AQI 指数）”右侧的三角形按钮，在打开的下拉菜单中选择【度量（总计）】|【平均值】。“行”功能区显示为“平均值（AQI 指数）”。单击工具栏上的“降序排序”按钮，或者单击 *Y* 轴“平均值 AQI 指数”后面的“降序排序”按钮，图表按“平均值 AQI 指数”字段降序排列。

（3）显示标记标签。单击工具栏中的“显示标记标签”按钮，在图表中显示当前工作表的标记标签，即显示各省会城市的“AQI 指数”的平均值。也可以在“标记”卡中选择“标签”，勾选“显示标记标签”复选框。

（4）编辑颜色。此时的图表虽然显示了标签并且按降序排列，但颜色单一，可以添加颜色区分空气质量。按住【Ctrl】键，拖动“行”功能区的“平均值（AQI 指数）”到“标记”卡中的“颜色”上，如图 5.46 所示。单击“标记”卡中的“颜色”，在打开的面板中单击“编辑颜色”按钮，如图 5.47 所示。在“编辑颜色”对话框中设置色板为“红色—蓝色发散”，设置渐变颜色为“4 阶”，“开始”为零，“结束”为 200，“中心”为 100，如图 5.48 所示，单击“确定”按钮。

图 5.46　添加颜色

图 5.47　单击“编辑颜色”按钮

1　http://www.zhb.gov.cn。

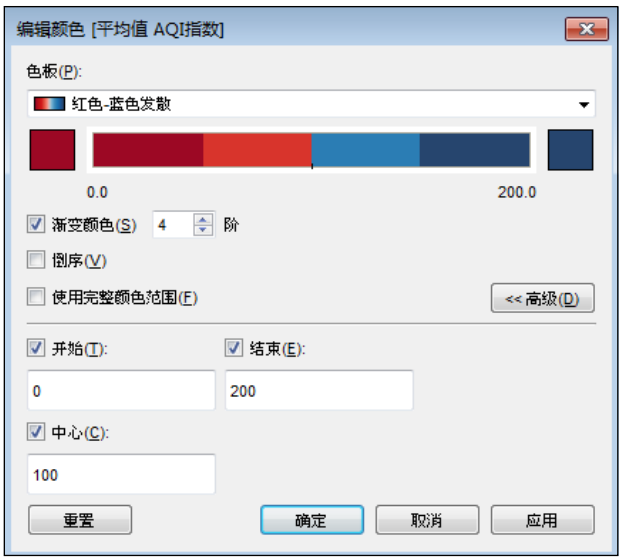


图 5.48　编辑简单条形图“AQI 均值排序”的颜色

注意，空气质量与 AQI 指数相关，一般选用的规则是：AQI 指数在 0～50 之间，空气质量是“优”；AQI 指数在 51～100 之间，空气质量是“良”；AQI 指数在 101～150 之间，空气质量是“轻度污染”；AQI 指数在 151～200 之间，空气质量是“中度污染”；AQI 指数在 201～300 之间，空气质量是“重度污染”；AQI 指数超过 300，空气质量是“严重污染”。渐变颜色设置为 4 阶，是因为平均空气质量从海口的 42.3 到哈尔滨的 177.7，分为 4 个阶段可以体现平均空气质量的“优”、“良”、“轻度污染”和“中度污染”四个等级差异。

最终效果如图 5.49 所示。该条形图显示了省会城市 2015 年 11 月共 30 天的 AQI 指数均值降序图。通过该图可以分析出，11 月份东北和华北地区省会城市的空气质量较差，其中东北三省的省会城市高居 AQI 指数均值前三名，石家庄、北京次之，这五个省会城市的 AQI 指数均值超过 150，平均空气质量是“轻度污染”。“福州”和“海口”两个省会城市的 AQI 指数均值最低，低于 50，平均空气质量是“优”。

案例 3：制作包含计算的堆叠条形图“各省会城市空气质量天数统计图”。

该条形图统计 2015 年 11 月各省会城市六种空气质量（“优”、“良”、“轻度污染”、“中度污染”、“重度污染”和“严重污染”）的天数，以方便读者查看各省会城市的空气质量情况。

（1）创建计算字段。单击“维度”中的“空气质量级别”字段后面的三角形按钮，在打开的下拉菜单中选择【创建】|【计算字段】，在打开的“计算字段”对话框中输入字段的名字“计数：空气质量级别”，然后输入公式“count（[空气质量级别]）”，如图 5.50 所示，单击“确定”按钮。

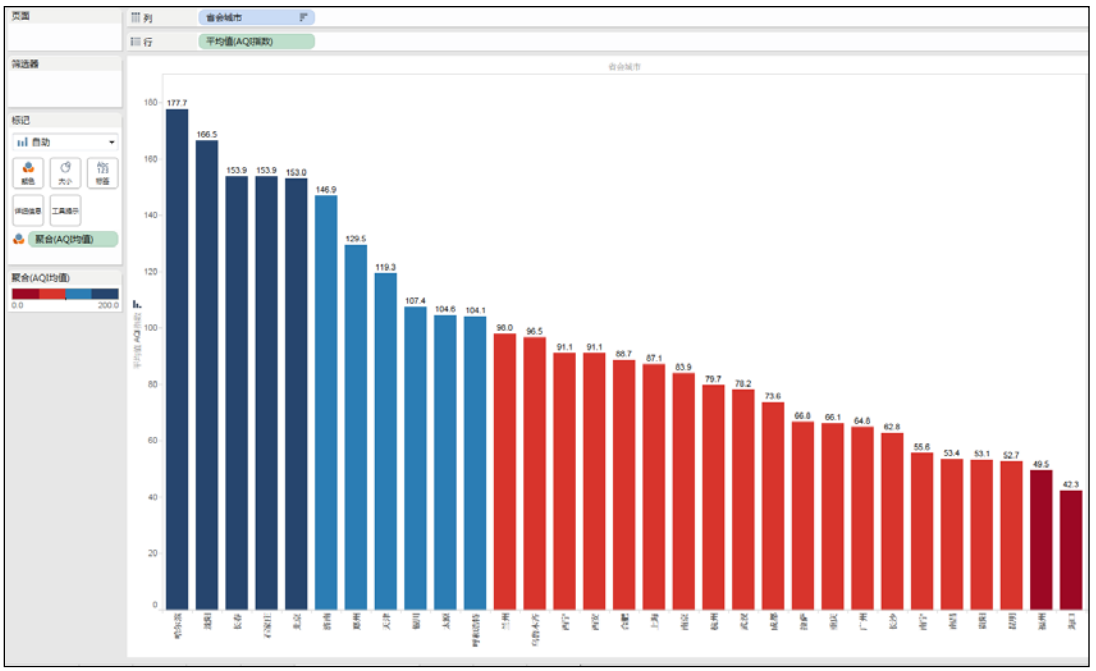


图 5.49　简单条形图“AQI 均值排序”

注意：公式中除字段名称外均为半角英文，特别注意英文括号。当左下角提示“计算有效”时表明公式正确，显示“计算包含错误”时说明公式有错误，单击“计算包含错误”后面的三角形图标可以查看公式错误的具体原因。

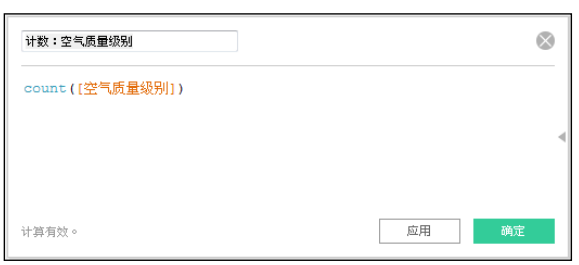


图 5.50　计算字段“计数：空气质量级别”

（2）制作堆叠条形图。同时选择“省会城市”、“空气质量级别”两个维度字段和“计数：空气质量级别”度量字段，展开“智能显示”，单击“堆叠条”选项。

（3）修改图例次序。默认情况下，“空气质量级别”图例显示的是“良”、“轻度污染”、“严重污染”、“优”、“中度污染”和“重度污染”，即默认情况下按中文拼音字母升序排序，如图 5.51 所示。但用户更关心按空气实际质量的级别排序，即按“优”、“良”、“轻度污染”、“中度污染”、“重度污染”和“严重污染”的序列排序。可以单击“空气质量级别”图例右上角的三角形按钮，在打开的

下拉菜单中选择【排序】选项，如图 5.52 所示。在弹出的对话框中选择“手动”单选框，在下面的文本框中设置排序，如图 5.53 所示。单击“确定”按钮。

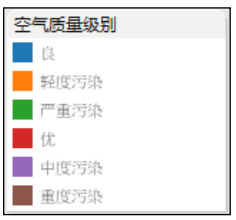


图 5.51　默认“空气质量级别”图例

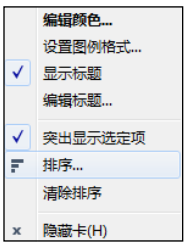


图 5.52　选择【排序】选项

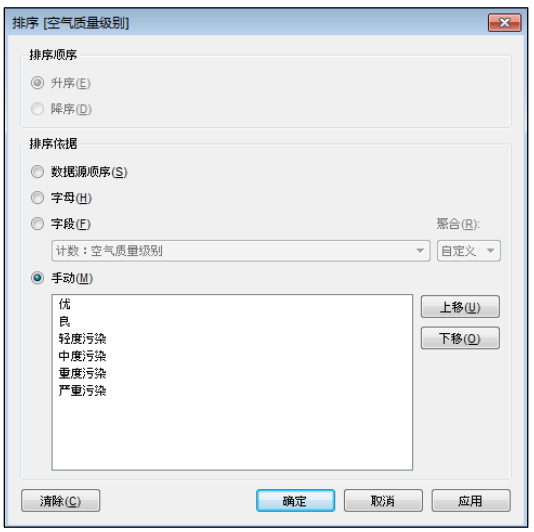


图 5.53　“排序”对话框

（4）显示标记标签。单击工具栏中的“显示标记标签”按钮，显示各省会城市“优”、“良”、“轻度污染”、“中度污染”、“重度污染”和“严重污染”六种空气质量的天数，如图 5.54 所示。

通过“各省会城市空气质量天数统计图”可以看出，绝大多数省会城市的空气质量以“优”和“良”为主，即红色和蓝色天数较多。其中，“海口”的空气质量以“优”的天数最多，高达 24 天，而“沈阳”、“哈尔滨”和“济南”三座城市却是“轻度污染”的天数最多，分别是 13 天、8 天和 7 天。虽然通过简单条形图“AQI 均值排序”可以看出哈尔滨的平均 AQI 质量最高，即平均空气质量最差，但“严重污染”天数最多的却是“长春”，共 6 天，然后是“北京”、“石家庄”和“哈尔滨”均为 4 天。

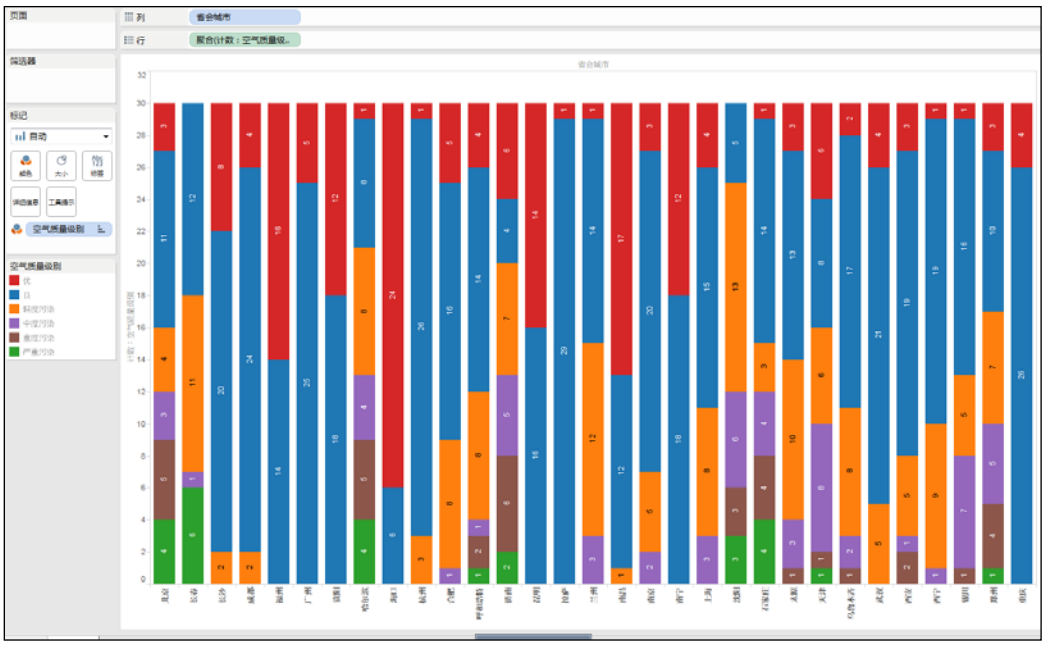


图 5.54　堆叠条形图“各省会城市空气质量天数统计图”

案例 4: 将其他图形转换为条形图“各省会城市空气质量百分比降序统计图”。

（1）制作文本表。该图统计 2015 年 11 月各省会城市六种空气质量（“优”、“良”、“轻度污染”、“中度污染”、“重度污染”和“严重污染”）天数的比例，以方便用户查看各省会城市的空气质量情况。使用“省会城市”、“空气质量级别”维度字段和“计数：空气质量级别”度量字段制作文本表，在“空气质量级别”中选择“优”，按住鼠标左键拖动到首行的位置。使用同样的方法将六种空气质量拖动到合适的位置，如图 5.55 所示。

（2）降序排列。按“空气质量级别”中“优”的天数降序排列，在“空气质量级别”中选择“优”，单击右侧的“降序排序”按钮，或者单击工具栏中的“降序排序”按钮，如图 5.56 所示。

列	省会城市		
行	空气质量级别		
空气质量级别	北京	长春	长沙
优	3		8
良	11	12	20
轻度污染	4	11	2
中度污染	3	1	
重度污染	5		
严重污染	4	6	

图 5.55　文本表“空气质量百分比统计”

列	省会城市		
行	空气质量级别		
空气质量级别	海口	南昌	福州
优	24	17	16
良	6	12	14
轻度污染		1	
中度污染			
重度污染			
严重污染			

图 5.56　按“优”降序排序

（3）转换图表类型。展开“智能显示”，单击“堆叠条”，将“文本表”图表转换为“堆叠条”。

（4）添加表计算。统计各省会城市空气质量的百分比。单击“行”功能区的“聚合（计数：空气质量级别）”字段右侧的三角形按钮，在打开的下拉菜单中选择【添加表计算】。在弹出的“表计算［计数：空气质量级别］”对话框中设置“计算类型”为“总额百分比”，设置“值汇总范围”为“空气质量级别”，如图 5.57 所示。单击“确定”按钮。

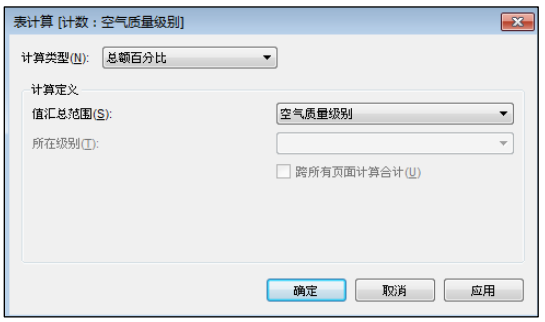


图 5.57　“表计算［计数：空气质量级别］”对话框

（5）显示标记标签。单击工具栏中的“显示标记标签”按钮。

（6）添加常量线。在“分析”选项卡中单击“常量线”，如图 5.58 所示，按住鼠标左键将其拖动到图表中，如图 5.59 所示，选择“表”，设置值为 0.5。可以根据用户个人兴趣设置常量线的值，也可以添加多根常量线，如用同样的方法再添加一根参考线，设置值为 0.25。

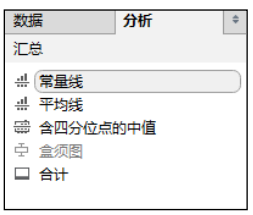


图 5.58　单击“常量线”

图 5.59　添加参考线到图表中

（7）设置百分比格式。图表中的百分比默认显示 2 位小数，略显数据凌乱，可将百分比设置为整数，缩短标签显示内容。单击“行”功能区的“聚合（计数：空气质量级别）”字段右侧的三角形按钮，在打开的下拉菜单中选择【设置格式】。在打开的对话框中选择“区”选项卡，单击“默认值”中“数字”右侧的三角形按钮，选择“百分比”的小数位数是“0”。

最终效果如图 5.60 所示，通过“各省会城市空气质量百分比降序统计图”，可以方便地查看空气质量是“优”的天数比例最多的城市是“海口”、“南昌”和“福州”，分别是 80%、57%和 53%。；“长春”和“沈阳”整个月份空气质量是“优”的天数比例均是零；“海口”、“福州”、“昆明”、“贵阳”、“南宁”、“广州”、“重庆”和“拉萨”这 8 座城市整个月份空气质量均为“优”和“良”，属于宜居城市。

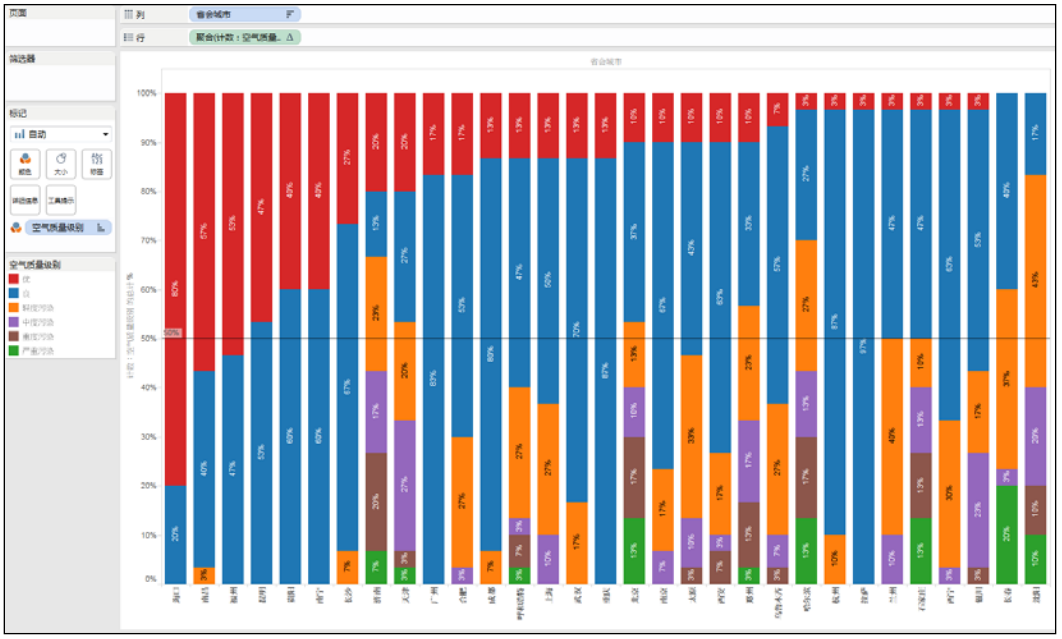


图 5.60　转换的条形图“各省会城市空气质量百分比降序统计图”

通过参考线可以看出，仅有“济南”、“天津”、“北京”、“郑州”、“哈尔滨”、“长春”和“沈阳”这七座城市整个月份超过一半的空气质量是“良”以下。

假设一个月里超过 25%的空气质量是“重度污染”或“严重污染”的城市是有必要继续改善空气质量的，则根据参考线可以看到，“济南”、“北京”、“哈尔滨”和“石家庄”均是需要改善空气质量的城市，因为这 4 座城市空气质量是“重度污染”或“严重污染”的比例分别是 27%、30%、30%和 27%。

5.6.5　线图

线图（也称折线图）是将数据视图中的各数据点连接起来。线图为直观显示一系列数值提供了一种简单的方法。线图非常适合显示数据随时间变化的趋势，或者预测未来的数值。折线图和条形图是在数据可视化中使用最多的两种图表类型。

案例 5: 制作简单线图“每天所有省会 AQI 均值图”。

该线图按日期统计所有省会 AQI 均值，方便读者查看全国各省会城市的整月空气质量趋势。

首先，连接数据“省会城市空气质量.xlsx”。然后，同时选择“日期”维度字段和“AQI 指数”度量字段，展开“智能显示”，单击“线（连续）图”。单击“列”功能区“年（日期）”字段后面的三角形按钮，在打开的下拉菜单中选择“日”，再单击工具栏中的“显示标记标签”，效果如图 5.61 所示。

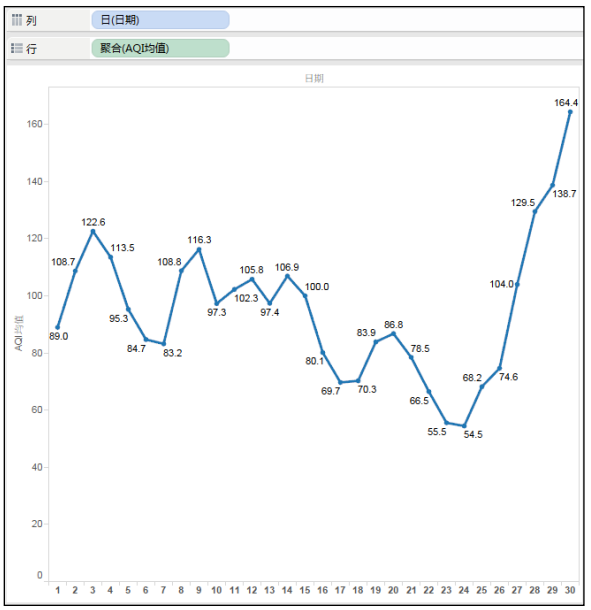


图 5.61　简单线图“每天所有省会 AQI 均值图”

最终效果视图显示出 2015 年 11 月全国所有省会城市 AQI 均值的情况，通过该图可以发现随着时间的推移，AQI 均值整体呈现升高的趋势，全国省会城市 AQI 均值最低的是 2015 年 11 月 24 日，均值是 54.5。11 月 24 日以后 AQI 均值逐日迅猛增加，最高到 164.4。

案例 6: 制作线图“重点省会城市 AQI 值图”。

为深入地分析 AQI 均值，结合前面的分析结果，可以对重点城市做进一步分析，这 7 座城市分别是“济南”、“天津”、“北京”、“郑州”、“哈尔滨”、“长春”和“沈阳”(选择这 7 座城市的原因是它们整个月份超过一半时间的空气质量在“良”以下)。

可以在“每天所有省会 AQI 均值图”的基础上编辑制作“重点省会城市 AQI 值图”。首先复制图表，然后进行编辑。

(1) 复制图表。用鼠标右键单击“每天所有省会 AQI 均值图”工作表标签，在弹出的快捷菜单中选择【复制工作表】，再单击标签栏中的“新建工作表”，用鼠标右键单击新建的工作表，在弹出的快捷菜单中选择【粘贴工作表】。操作后该工作表名称是“每天所有省会 AQI 均值图 (2)”。

(2) 重命名图表。用鼠标右键单击“每天所有省会 AQI 均值图 (2)”工作表标签，在弹出的快捷菜单中选择【重命名工作表】，输入新名称“重点省会城市 AQI 值图”。

(3) 筛选城市。在“维度”区选择“省会城市”字段，拖动到“标记”卡中的“颜色”上，在弹出的警告框中单击“筛选后添加”。在打开的“筛选器[省会城市]”对话框中勾选“济南”、“天津”、“北京”、“郑州”、“哈尔滨”、“长春”和“沈阳”，如图 5.62 所示，单击“确定”按钮。

(4) 编辑颜色。筛选的 7 座城市的颜色可能相近，不方便区分，可以为 7 座城市重新设置颜色。单击“标记”卡中的“颜色”，再单击“编辑颜色”按钮，在打开的“编辑颜色[省会城市]”对话框

中设置“选择调色板”为“Tableau10”，如图 5.63 所示，单击“分配调色板”按钮，可以根据自己的喜好为每座城市设置颜色。

（5）取消显示标记标签。这 7 座城市共 210 个标记标签的显示略显凌乱，单击工具栏中的“显示标记标签”按钮，取消标记标签的显示。

（6）编辑工具提示信息。单击“标记”卡中的“工具提示”，在打开的“编辑工具提示”对话框中输入提示信息，如图 5.64 所示。注意，尖括号中的内容是字段名称或计算字段等，不建议修改，其他内容可以根据自己的喜好修改。

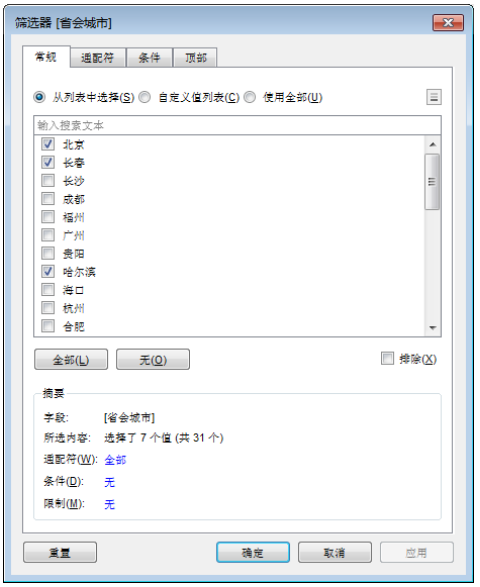


图 5.62　“筛选器[省会城市]”对话框

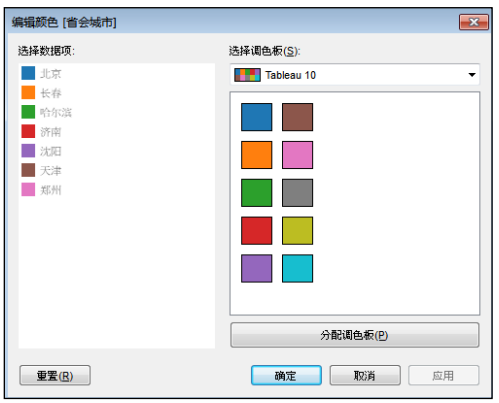


图 5.63　“编辑颜色[省会城市]”对话框

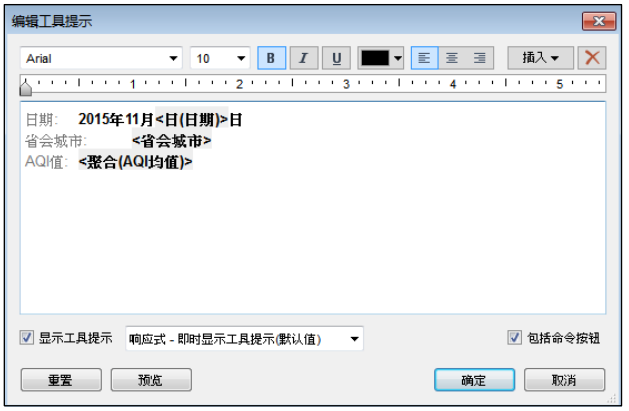


图 5.64 “编辑工具提示”对话框

（7）设置格式。AQI 数值显示时带有 1 位小数，而实际上 AQI 值均为整数。单击“行”功能区的“聚合（AQI 均值）”字段右侧的三角形按钮，在打开的下拉菜单中选择【设置格式】。在打开的对话框中选择“区”选项卡，单击“默认值”中“数字”右侧的三角形按钮，在打开的下拉菜单中选择“数字（自定义）”的小数位数是“0”。

最终效果如图 5.65 所示。查看这 7 座城市的 AQI 值，可以发现“北京”、“济南”、“天津”和“郑州”这 4 座城市的 AQI 值随着时间逐步升高，而“长春”、“哈尔滨”和“沈阳”正好相反，均有下降的趋势。

图 5.65 线图“重点省会城市 AQI 值图”

案例 7: 制作线图“重点城市首要污染物图”。

不同的省会城市首要污染物是不同的，该线图按日期统计重点省会城市 AQI 值与首要污染物的关系，以查看 AQI 值高的重点城市以哪种首要污染物为主（重点城市是“济南”、“天津”、“北京”、“郑州”、“哈尔滨”、“长春”和“沈阳”）。

（1）制作线图。同时选择“日期”、“省会城市”、“首要污染物”和“AQI 指数”字段，展开“智能显示”，选择“线（连续）图”，将“列”功能区展开为“日（日期）”，筛选 7 座重点城市，如图 5.66 所示。

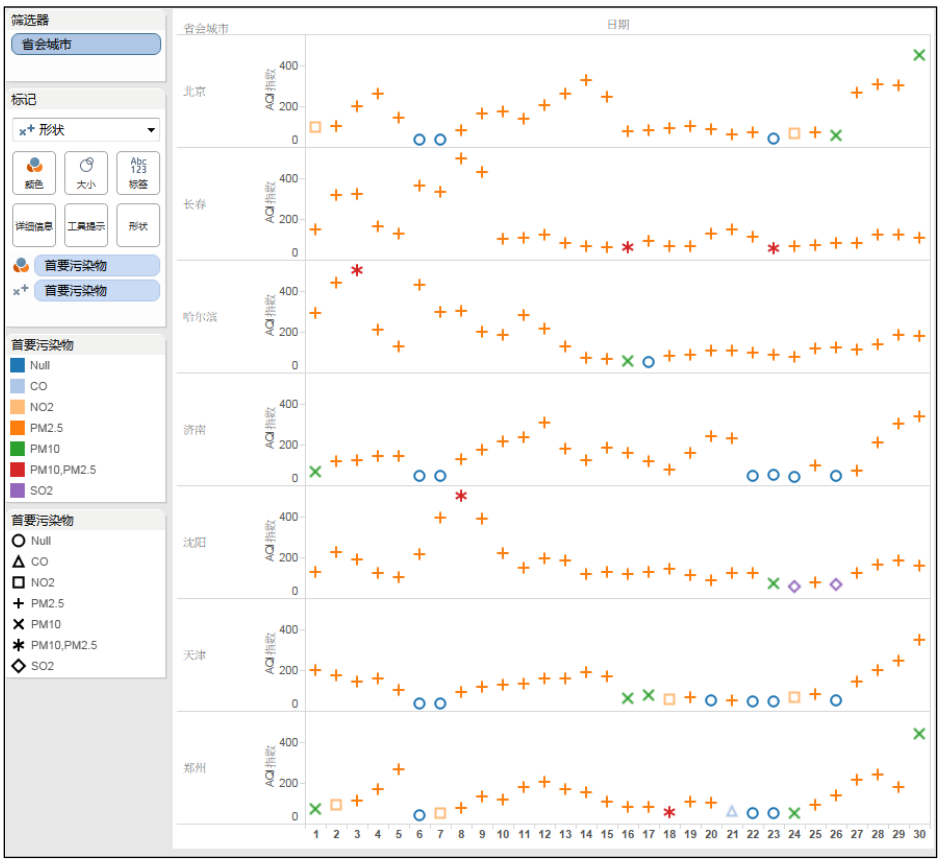


图 5.66　线图　“重点城市首要污染物图”

（2）编辑形状。在“标记”卡中选择“形状”，然后将“维度”中的“首要污染物”字段拖动到“标记”卡的“形状”上。

从最终效果可以看到，这 7 座重点城市的首要污染物是 PM2.5（“+”号最多）。因此，改善空气质量等级的首要任务是去除 PM2.5，这 7 座重点城市的市民出行要购买防 PM2.5 的口罩。其中，“天津”、“济南”和“长春”这 3 座城市空气质量最差的天气首要污染物是 PM2.5,“北京”和“郑州”

空气质量最差的天气首要污染物是 PM10，“哈尔滨”和“沈阳”空气质量最差的天气首要污染物是 PM2.5 和 PM10。

案例 8：可视化图表类型转换。

线图可以直接制作成条形图，也可通过已制作的其他图表类型转换。但要注意，并不是任何一种图表类型均可转换为其他图表类型。

（1）制作一个并排条图。将“空气质量级别”和“省会城市”两个维度字段拖到“列”功能区（注意顺序），再将“计数：空气质量级别”度量字段拖到“行”功能区。将“省会城市”维度字段拖动到“颜色”上，并筛选“济南”、“天津”、“北京”、“郑州”、“哈尔滨”、“长春”和“沈阳”这 7 座城市（注意，这 7 座城市整个月份超过一半时间的空气质量在“良”以下，筛选方法参见图 5.62）。

（2）编辑并排条图。空气质量级别按“优”、“良”、“轻度污染”、“中度污染”、“重度污染”和“严重污染”排序显示（方法参见图 5.53），显示标记标签。

（3）图表适应整个视图。单击工具栏中的“适合选择器”下拉按钮，在打开的下拉列表中选择“整个视图”（参见表 5.1）。

（4）图表类型转换。展开“智能显示”，线图的颜色都是虚的，说明并排条图不能转换为线图，但可以转换为其他有颜色的图表类型，如树地图或气泡填充图等。

最终效果如图 5.67 所示，可视化图表类型可以互相转换，但并不一定转换为某种特定的类型，是否可以转换与可视化图表的维度个数和度量个数相关。

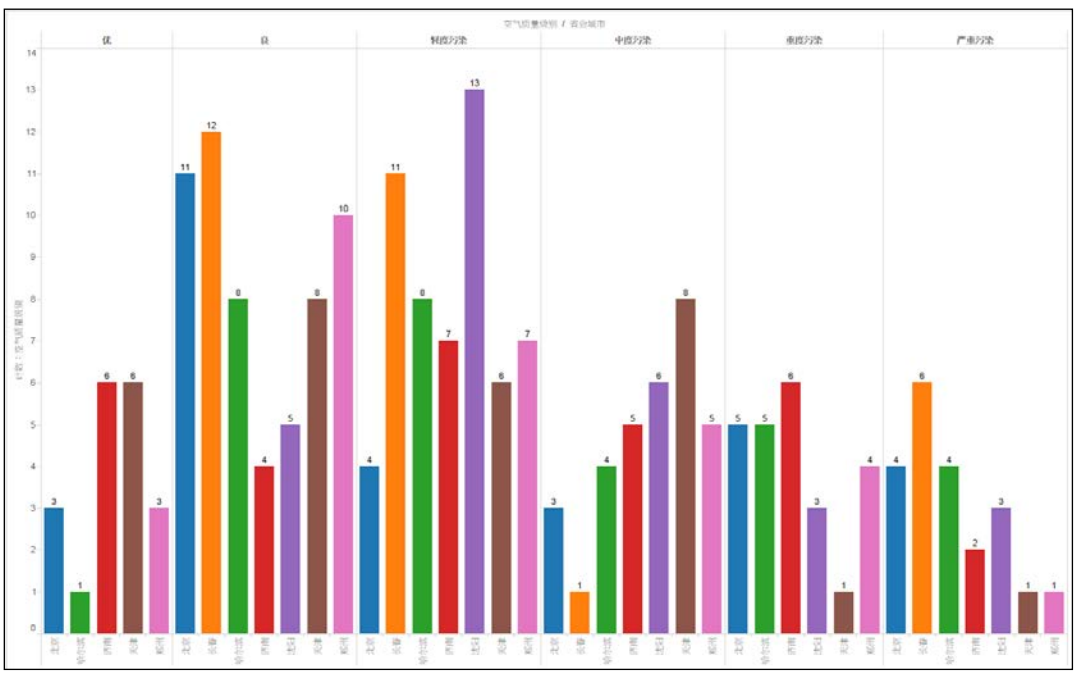


图 5.67　并排条图“7 座重点城市”

5.6.6　地图

地理地图简称为地图，是一种非常有用的图表类型，适合呈现包含地理位置信息字段的图表。地理位置信息可以是经度、纬度、城市、国家/地区、县、省/市/自治区和邮政编码等。只有包含地理位置信息的数据才可以制作地图，但并不是包含地理位置信息就必须使用地图展示数据，有时候条形图和线图也是很好的选择，特别是有聚合函数或分组的时候。

地图分为符号地图和填充地图（也称为热力地图）两种。符号地图是为地图上的每个位置显示一个标记，标记可以是圆形、正方形或饼形等形状。填充地图是根据数值大小用颜色填充地理区域的多边形。例如，可以根据各个省份 GDP 值用颜色填充省份的多边形区域。

案例 9：制作符号地图“AQI 均值地图”。

（1）连接数据。数据源连接到文件“省会城市空气质量.xlsx”。

（2）新建计算字段。单击“度量”中的“AQI 指数”字段后面的三角形按钮，在打开的下拉菜单中选择【创建】|【计算字段】选项，在打开的“计算字段”对话框中输入字段的名称“AQI 均值”，输入公式“AVG（[AQI 指数]）”，单击“确定”按钮。

（3）转换字段角色。“省会城市”字段在“维度”区，默认为文本型数据，为制作数据地图，需要转换该字段的角色。用鼠标右键单击“省会城市”字段，在打开的快捷菜单中选择【地理角色】|【城市】选项。Tableau 可进行地理编码的信息类型如表 5.3 所示。

表 5.3　Tableau 可进行地理编码的信息类型

地 理 角 色	说　　明
地区代码	美国地区代码；仅限数字。例如，206、650、415
CBSA/MSA	美国基于内核的统计区域或都市统计区域。例如，Dallas-Fort Worth-Arlington、TX
城市	全世界的城市名称。例如，Seattle、北京、东京
国会选区	美国国会选区名称。边界由每个州的重划选区委员会提供的数据确定。例如，1st District、2、District 3、4th
国家/地区	全球国家/地区。包括名称、FIPS 10、2 字符（ISO 3166-1）或 3 字符（ISO 3166-1）。例如，AF、CD、Japan、Australia、BH、AFG、UKR
County	所选国家/地区的二级行政区域。例如，美国的郡/县、法国的 départements、德国的 kriese 等
纬度	以十进制度数为单位的纬度，只能用于数字字段
经度	以十进制度数为单位的经度，只能用于数字字段
省/市/自治区	全世界范围的州/省/市/自治区，以及其他一级行政区域。可以用英语、法语、德语、西班牙语、葡萄牙语、日语、韩语和简体中文表示
邮政编码	美国、法国、德国、英国、加拿大、澳大利亚和新西兰的邮政编码。例如，法国邮政编码 75000、英国邮政编码 SO16 3ZG

（4）制作符号地图。按住【Ctrl】键，同时选择“省会城市”维度字段和“AQI 均值”度量字段，展开“智能显示”，选择符号地图。

（5）美化数据地图。在“维度”区选择“省会城市”字段，拖动到“标记”卡中的“颜色”上，

各个城市的圆形标记呈现出不同的颜色。单击“标记”卡中的“颜色”，设置黑色边界、灰色光环。单击“标记”卡中的“大小”，增大标记到合适的大小。单击“标记”卡中的三角形下拉按钮，在下拉菜单中选择“饼图”，如图 5.68 所示。单击“标记”卡中的“标签”，勾选“显示标记标签”和“允许标签覆盖其他标记”复选框，如图 5.69 所示。

图 5.68　设置“饼图”

图 5.69　设置“标签”

最终效果如图 5.70 所示，该图显示了城市和 AQI 均值的关系。各省会城市按颜色区分，显示在相应的地理位置，AQI 均值显示在圆圈内。分析地图发现，东北和华北地区在 2015 年 11 月份的空气质量最差，其中东北三省的省会城市的空气质量高居 AQI 指数前三名，石家庄、北京次之。沿海省会城市的 AQI 均值低，如上海、福州、广州、海口和南宁，越接近内陆城市 AQI 值越高。

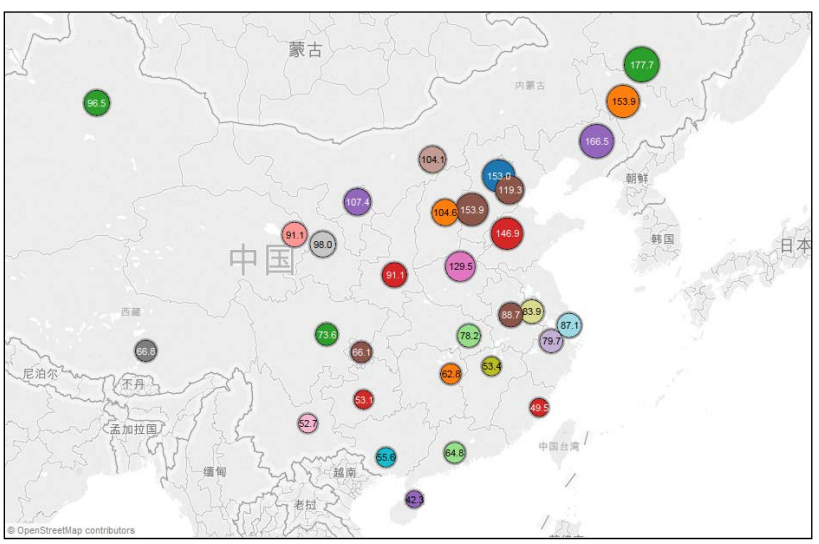


图 5.70　符号地图“AQI 均值地图”

（6）未标记地理位置的处理。如果存在某个城市无法匹配的情况，可以在匹配位置查询匹配城市，也可以用网络查找该城市的经纬度，然后输入到“匹配位置”，如“海口”的纬度和经度分别是 20.02 和 110.35，如图 5.71 所示。

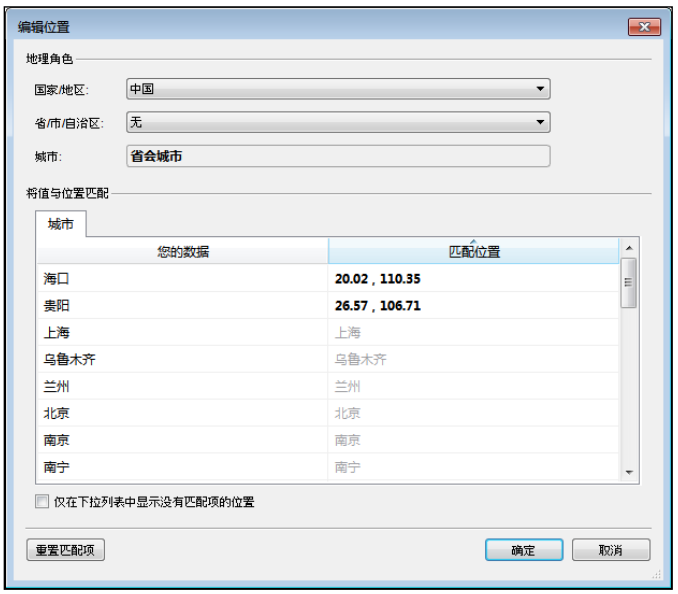


图 5.71　“编辑位置”对话框

注意，地图可以匹配“北京”和“吉林”，即可以匹配到地图上该省份或城市的经纬度，但不匹配“北京市”和“吉林省”，若数据是后一种情况，需要使用 OpenRefine 清理数据后使用。

地图可帮助用户快速查找位置和分析全球范围内的数据。通过地图搜索及平滑平移和缩放体验，可以轻松地浏览地图中的数据。

单击 图标并在搜索框中输入位置名称，地图将平移和缩放到该位置。在搜索框中可以输入州名、国家/地区、省/市/自治区、县、城市或邮政编码等。

若要平移，可在地图视图中的任何位置单击并拖动。

若要缩放，可以双击地图上的某个区域，或者使用视图左上角视图工具栏中的缩放（ + 或 − ）控件。

若要在浏览后返回到地图的初始视图，可以单击视图工具栏中的 按钮。

案例 10: 填充地图“利润地图”。

本地图是 5.4.4 小节“思考”中的“利润地图”。按国家计算利润总和，并用颜色填充，颜色越绿表示盈利越多，颜色越红表示亏损越多，参见图 5.20。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。选择“订单”表并拖动到数据区。

（2）制作填充地图。在“维度”区选择“国家/地区”字段，按住【Ctrl】键的同时在“度量”

区选择“利润”字段。在“智能显示”中单击“填充地图”。

（3）设置地图选项。创建地图视图时，有多个选项可帮助编辑地图的外观。单击【地图】|【地图选项】选项后打开“地图选项”面板，可以修改地图背景、设置隐藏和显示层（例如街道名称和国家/地区边界）、添加数据层等，如图 5.72 和图 5.73 所示。

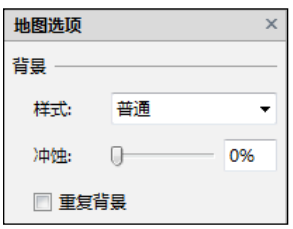


图 5.72　修改地图背景

图 5.73　修改地图层

地图背景样式共有三种，“普通”背景效果如图 5.74 所示，“浅”背景效果如图 5.75 所示，“黑色”背景效果如图 5.76 所示。

图 5.74 地图“普通”背景

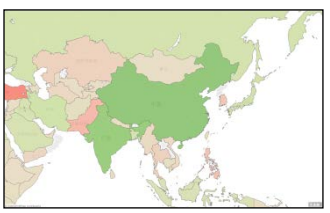


图 5.75 地图“浅”背景

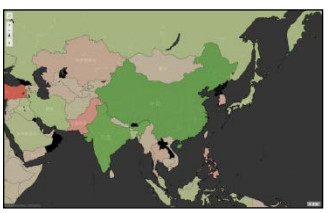


图 5.76 地图“黑色”背景

可以自定义用户与地图的交互方式，如希望限制用户与地图的某些交互方式。常见的交互方式有以下两种。

一是隐藏地图搜索。单击【地图】|【显示地图搜索】选项可以隐藏地图搜索图标，使用户无法在地图视图中搜索位置。

二是隐藏视图工具栏。在地图中单击鼠标右键，在打开的快捷菜单中选择【隐藏视图工具栏】选项隐藏视图工具栏，用户将无法将地图锁定到适当的位置，或将地图自动缩放以显示所有的数据。

注意：隐藏视图工具栏后，用户仍然可以使用键盘快捷方式来缩放视图、进行平移及选择标记。

5.6.7　饼图

饼图特别适合显示比例，尤其适合分组较少的情况下使用（如空气质量分为 6 个等级），且每个分组的百分比不能太大或太少。

下面使用 5.6.4 小节的数据“省会城市空气质量.xlsx”制作饼图“北京空气质量等级饼图”，显示 2015 年 11 月北京各种空气质量级别的累计天数百分比；利用饼图与地图结合制作数据化图表“各省会城市空气质量等级百分比地图”，在地图上显示各省会城市空气质量等级百分比数据。

案例 11：制作饼图“北京空气质量饼图”。

（1）连接数据。数据源连接到文件“省会城市空气质量.xlsx”。

（2）制作饼图。将“省会城市”维度字段拖到“列”功能区，将“空气质量级别”维度字段拖到“行”功能区，再将“空气质量级别”维度字段拖到“行”功能区。单击“行”功能区第二个“空气质量级别”字段后面的三角形按钮，在打开的下拉菜单中选择【度量】|【计数】。展开“智能显示”，选择饼图。筛选“北京”，显示标记标签“空气质量级别”和“百分比：空气质量级别”。

最终效果如图 5.77 所示。为了更好地显示空气质量天数百分比，我们希望的效果如图 5.78 所示。

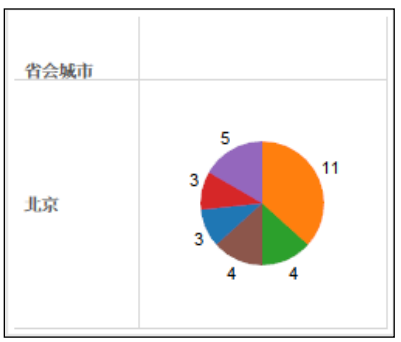


图 5.77　统计北京空气质量等级天数

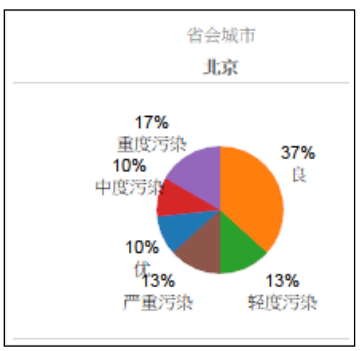


图 5.78　统计北京空气质量等级百分比

（3）新建计算字段。单击“维度”中的“空气质量级别”字段后面的三角形按钮，在打开的下拉菜单中选择【创建】|【计算字段】选项，在打开的“计算字段”对话框中输入字段的名称“百分比：空气质量级别”，输入公式“count（[空气质量级别]）/30”，单击“确定”按钮。

（4）美化饼图。将“百分比：空气质量级别”字段拖动到“标记”卡中的“标签”上，按空气质量级别统计的比例（小数）将显示在饼图上，单击“标记”卡中的“聚合（百分比：空气质量级别）”后面的三角形按钮，在打开的下拉菜单中选择【设置格式】，然后在打开的对话框中选择“区”选项卡，单击“默认值”中“数字”右侧的三角形按钮，选择“百分比”的小数位数是“0”。

得到如图 5.78 所示的效果，该图显示了 2015 年 11 月北京空气质量以“良”为主，占 37%，其次是“重度污染”天气，占 17%。

案例 12：制作饼图“天津空气质量饼图”和“上海空气质量饼图”。

使用上述方法制作“天津空气质量饼图”和“上海空气质量饼图”，如图 5.79 和图 5.80 所示。

如果标签过小或未看到标签，可以放大饼图以确保大多数单独的标签均可见。放大饼图的快捷键是【Ctrl】+【Shift】+【B】。若标签过大，可以缩小饼图，快捷键是【Ctrl】+【B】。若标签显示不全，可以单击“标记”卡中的“标签”，勾选“允许标签覆盖其他标记”复选框即可。

注意图 5.80 和图 5.81 的差异，图 5.80 中的部分标签依旧是有覆盖的，而图 5.81 根据比例值调整了数据在饼图中的排序（方法见 5.6.4 小节的图 5.53“排序”对话框），完美地解决了这个问题。

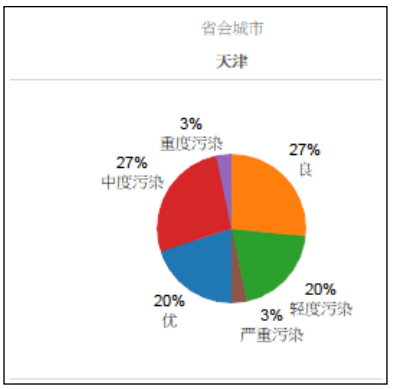


图 5.79　统计天津空气质量等级百分比

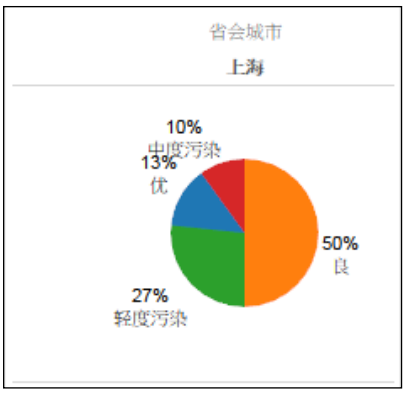


图 5.80　统计上海空气质量等级百分比

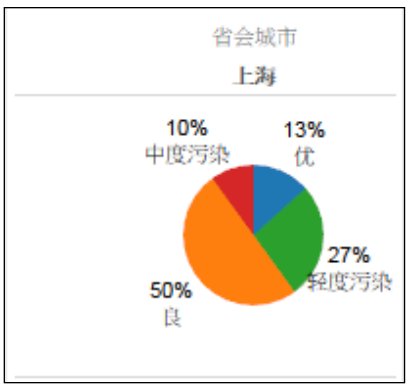


图 5.81　修改后的统计上海空气质量等级百分比

案例 13：饼图与地图结合制作“各省会城市空气质量等级百分比地图”。

从前面的学习中可以发现，饼图和堆叠条形图均适合显示比例，饼图更适合分组较少的情况，堆叠条形图对分组没有具体要求，但有时候更适合使用饼图与地理位置相关的属性共同呈现图表信息（如图 5.82 所示），该图在地图的基础上增加了饼图来展现各省会城市空气质量等级百分比。

（1）复制工作表。复制“AQI 均值地图”到新建的工作表，重命名为“各省会城市空气质量等级百分比地图”。

（2）编辑图表。将“空气质量级别”维度字段拖动到“标记”卡中的“颜色”上，每个省会城市均显示一个饼图。但标签显示的是 AQI 均值，且标签互相覆盖，过于紧密的标签并不美观。单击“标记”卡中的“标签”，取消“显示标记标签”和“允许标签覆盖其他标记”复选框的勾选。将“百分比：空气质量级别”字段拖动到“标记”卡中的“标签”上。单击“标记”卡中的“聚合（百分比：空气质量级别）”后面的三角形按钮，在打开的下拉菜单中选择【设置格式】选项，然后在打开的对话框中选择“区”选项卡，单击“默认值”中“数字”右侧的三角形按钮，选择“百分比”的小数位数是“0”。

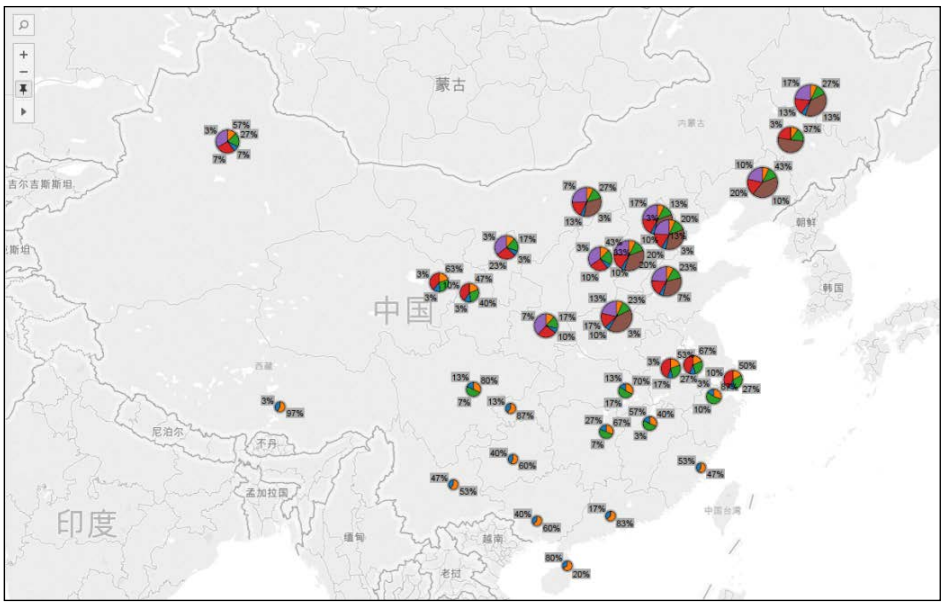


图 5.82　饼图与地图结合制作的“各省会城市空气质量等级百分比地图”

虽然 Tableau 从不使用饼图作为自动标记类型，但可以在“标记”卡的下拉菜单中选择“饼图”。

查看图 5.82 所示饼图的大小可以了解 AQI 均值，通过饼图可以清楚地分析出东北和华北空气质量等级为“严重污染”的天数比例最高的是“长春”，高达 20%，即 2015 年 11 月份“长春”有 20%的天数空气质量等级为“严重污染”，紧随其后的是“石家庄”、“北京”和“哈尔滨”，均有 13%的天数空气质量等级为“严重污染”。当然也可以通过堆叠条图“省会城市空气质量百分比降序统计图”分析得出类似信息。

5.6.8　树地图

树地图是一种相对简单的数据可视化形式，通过具有视觉吸引力的矩形块呈现信息。树地图是在嵌套的矩形块中显示数据。可使用维度定义树地图的结构，使用度量定义各个矩形块的大小和颜色。

树地图包含两个度量，一个度量控制大小，另一个度量控制颜色。树地图可以包含任意数量的维度，但只能通过一个维度控制颜色，实现视图的多样性。其他维度只能用于增加视图中矩形块的数量。

案例 14：使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”，制作树地图“利润树地图”，显示各子类产品的聚合总利润。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）制作树地图。将“子类别”维度字段拖到“列”功能区，将“利润”度量字段拖到“行”

功能区。此时将默认生成一个条形图（当“列”功能区上有一个维度且“行”功能区上有一个度量时的默认图表类型）。展开“智能显示”，选择树地图。将“销售额”字段拖到“标记”卡中的“大小”上，如图 5.83 所示。

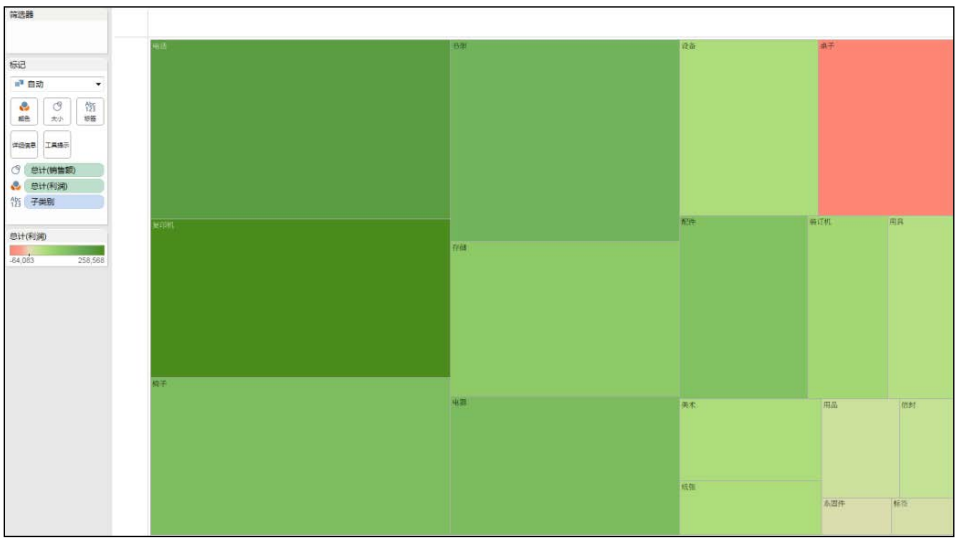


图 5.83　利润树地图

在最终效果图中，“利润”决定矩形块的颜色，“销售额”决定矩形块面积的大小，值越大，矩形块面积越大。通过矩形块大小可以快速看出“电话”的销售额总计最高，“标签”的销售额总计最低。销售额和利润的关系并不是呈现正比的规则，通过颜色可以看到“复印机”利润总计最高，“桌子”利润总计是负数，虽然“桌子”的销售额总计并不是最差的，但利润却是最低的。

5.6.9　填充气泡图

使用填充气泡图可以在一组圆中显示数据。其中，维度定义各个气泡，度量定义各个圆的大小和颜色。

案例 15：使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”，制作填充气泡图“装运成本填充气泡图”，显示各子类产品的聚合总装运成本。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）制作填充气泡图。选择“类别”、“子类别”维度字段和“装运成本”度量字段，展开“智能显示”，选择填充气泡图。

（3）美化填充气泡图。标签中包含的“类别”和“子类别”文字大小一样，不方便区分。可以单击“标记”卡中的“标签”按钮，在打开的面板中单击 “文本”文本框后面的按钮，如图 5.84 所示。在打开的“编辑标签”对话框中将“<类别>”的字号修改为“12”，如图 5.85 所示。

图 5.84　设置标签外观的文本格式

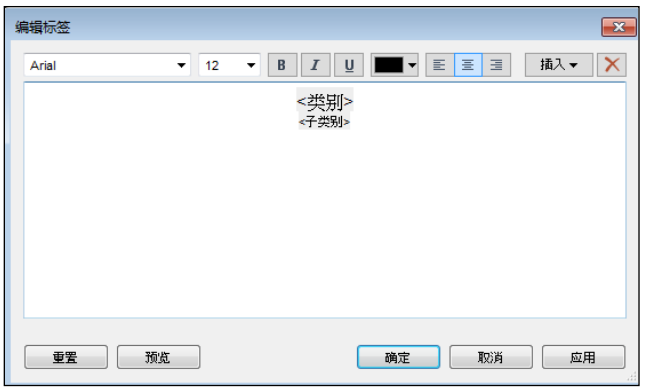


图 5.85　编辑标签中的字号

最终效果如图 5.86 所示，各子类别的装运成本总计决定气泡的大小，“类别”决定气泡的颜色，绿色表示“家具”类，蓝色表示“办公用品”类，橙色表示“技术”类产品。可以发现，“电话”类产品的装运成本最高，“标签”类产品的装运成本最低。

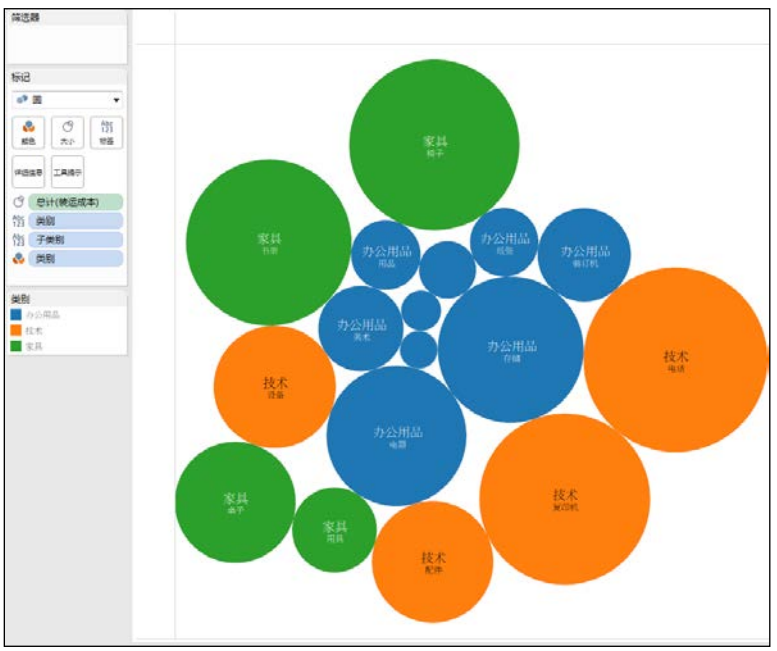


图 5.86　装运成本填充气泡图

5.6.10　甘特图

甘特图（也称甘特条形图）适合查看日期、项目计划或不同定量变量之间的关系。甘特图中每

个标记的长度都与“标记”卡的“大小”上放置的度量成比例。

案例 16：使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”，制作甘特图“订发时间差甘特图”，显示订购日期和发货日期的间隔天数。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）计算订发时间差字段。用鼠标右键单击“数据”窗格中的空白处，在打开的快捷菜单中选择【创建计算字段】选项。在打开的“计算字段”对话框中输入字段的名称“订发时间差”，输入公式“DATEDIFF（'day',[订购日期],[装运日期]）”，单击“确定”按钮。该公式计算“订购日期”和“装运日期”字段值的差，用“天数”作为计量单位。

（3）制作甘特图。选择“订购日期”和“邮寄方式”维度字段，展开“智能显示”，选择甘特图。在“列”功能区上，单击“订购日期”下拉按钮，然后在下拉列表中选择“周数”。将“子类别”拖到“行”功能区，并放到“邮寄方式”的左边（注意顺序）。

（4）美化甘特图。根据订购日期和发货日期之间的间隔天数来确定标记的大小。将“订发时间差”维度字段拖到“标记”卡中的“大小”上。该字段默认聚合为“总计”，单击“标记”卡上的“总计（订发时间差）”字段后的三角形按钮，在打开的下拉菜单中选择【度量（总计）】|【平均值】。将“邮寄方式”字段拖到“标记”卡中的“颜色”上。

（5）筛选时间。制作完成的甘特图中标记又多又密，其实可以仅显示筛选的部分时间。按住【Ctrl】键并将“周（订购日期）”字段从“列”功能区拖到“筛选器”功能区（注意，必须按住【Ctrl】键，否则“周（订购日期）”字段将移除“列”功能区）。在打开的“筛选器字段[订购日期]”对话框中选择“日期范围”，然后单击“下一步”按钮。在打开的“筛选器[订购日期]”对话框中设置日期范围是“2014/1/1”至“2014/3/31”，如图 5.87 所示，然后单击“确定”按钮（可以使用滑块确定日期，也可以直接在“日期”文本框中输入所需的数字或使用日历选择日期）。

图 5.87 “筛选器[订购日期]”对话框

最终效果如图 5.88 所示，可视化图表按“子类别”分类显示订购时间与发货时间之间的滞后天数，可以看出哪些邮寄方式更容易有较长的滞后时间、滞后时间是否因类别有异，以及滞后时间在一段时间内是否一致等具体问题。

图 5.88　订发时间差甘特图

5.6.11　散点图

散点图可以直观地显示数字变量之间的关系。在“列”功能区和“行”功能区上分别放置至少一个度量来创建散点图。如果这些功能区同时包含维度和度量，则度量将被设置为最内层字段，这意味着度量始终位于同样放置在这些功能区上的任何维度的右侧。散点图可以包含零个或几个维度，但至少包含两个、最多四个度量字段。

案例 17：使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”，制作散点图“利润与销售额散点图”，显示利润和销售额之间的关系及趋势线。趋势线可以提供利润和销售额两个字段数值之间关系的统计定义。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）制作散点图。选择“利润”和“销售额”度量字段，展开“智能显示”，选择散点图。Tableau 默认将两个度量聚合为总计。将“类别”维度字段拖到“标记”卡中的“颜色”上。数据分成了三种颜色标记。标记增加了很多，标记的数量等于数据源中不重复的地区数和类别数的乘积。

（3）添加趋势线。将“分析”窗格中的“趋势线”模型拖到散点图上，选择“线性”趋势线，如图 5.89 所示。

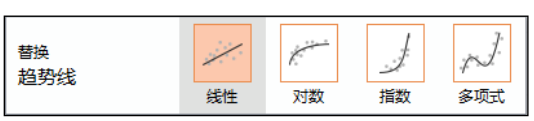


图 5.89　选择“线性”趋势线

（4）移除置信区间。Tableau 默认为每种颜色的标记添加趋势线（包含三条线），散点图共九条线。用鼠标右键单击散点图的空白处，在打开的快捷菜单中选择【趋势线】|【编辑趋势线】选项。然后在打开的“趋势线选项”对话框中，去掉“显示置信区间”复选框的勾选，如图 5.90 所示，单击“确定”按钮。

最终效果如图 5.91 所示，将鼠标指针悬停在线性趋势线上可查看有关用于创建该线的模型的统计信息“利润=.125556*销售额+3306.39,P 值:<.0001”。还可以尝试制作“对数”、“指数”和“多项式”型趋势线。如图 5.92 所示是多项式趋势线。

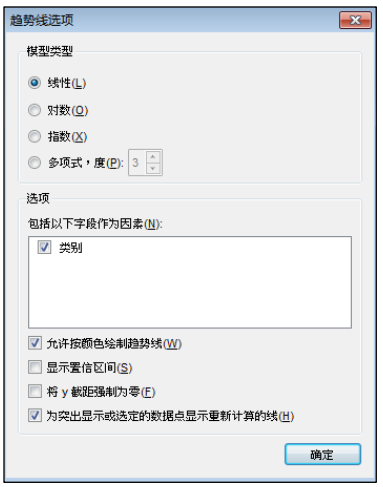


图 5.90 “趋势线选项”对话框

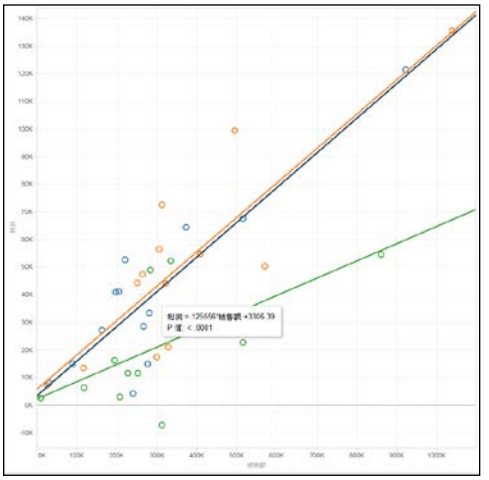


图 5.91 利润与销售额散点图（线性趋势线）

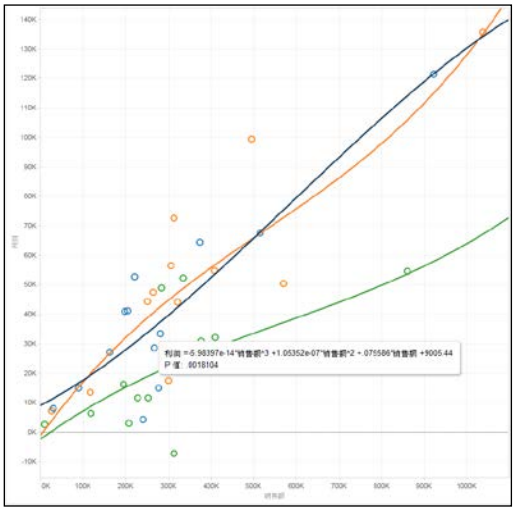


图 5.92 利润与销售额散点图（多项式趋势线）

（5）模型的显著性比较。图 5.91 和图 5.92 分别添加了不同种类的趋势线，通过查看 P 值（显著性）确定模型的拟合优度，即分析模型预测的质量。用鼠标右键单击散点图的空白处，在打开的快捷菜单中选择【趋势线】|【描述趋势模型】选项。如图 5.93 所示是利润与销售额散点图（线性趋势线）的描述趋势模型。

等于或小于 0.05 的 P 值是正常的。P 值（显著性）越小，模型的显著性就越高。通过分析可以发现，本案例中线性趋势线好于多项式趋势线。

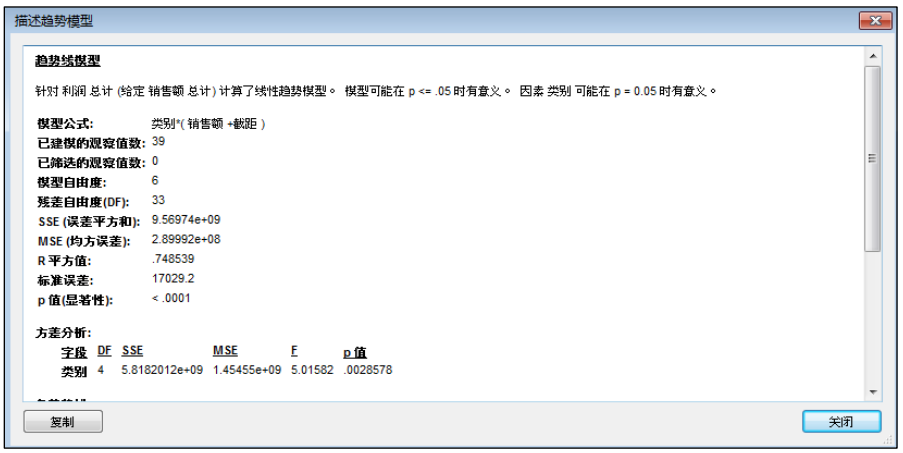


图 5.93　利润与销售额散点图（线性趋势线）的描述趋势模型

5.6.12　双组合图和面积图

面积图又称区域图，强调数量随日期变化的程度，也可用于引起用户对总值趋势的注意。面积图包含 1 个日期和至少 1 个度量字段。

双组合图是用两种可视化图表类型呈现信息的方法，包含 1 个日期和至少 2 个度量字段。

下面使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”，制作双组合图“利润与销售额双组合图”和面积图“利润与销售额面积图”，显示订购日期、类别、利润和销售额之间的关系。

案例 18：制作双组合图“利润与销售额双组合图”。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）制作双组合图。同时选择“订购日期”和“类别”维度字段，“销售额”和“利润”度量字段，展开“智能显示”，选择双组合图。Tableau 默认将“订购日期”按年份聚合，并创建具有年份标签的列标题“年(订购日期)”，用折线图和面积图显示“销售额”和“利润”。双组合图显示了 2011～2014 年每年的“销售额”和“利润”总计，共 4 个数据，可视化图表稀疏。

（3）美化双组合图。单击视图“列”功能区“年（订购日期）”字段的三角形按钮，在打开的下拉菜单中选择【月　2015 年 5 月】，视图比原来的视图更加详细，可查看 4 年内“销售额”和“利润”的连续范围。如图 5.94 和图 5.95 所示分别显示了整个图表的左上角部分和右上角部分。

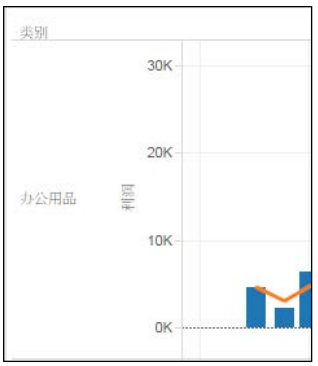


图 5.94 利润图

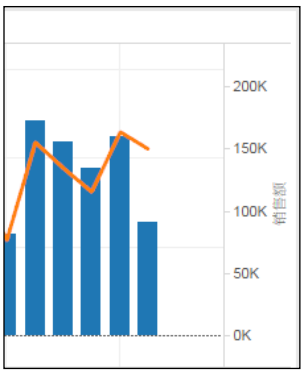


图 5.95 销售额图

（4）编辑轴。分析图 5.94 和图 5.95，发现利润和销售额数值相差不多，这从理论上是不太可能的（一般情况下，利润要远远小于销售额），仔细查看发现两幅图 *Y* 轴的刻度是不一样的。利润的刻度间隔是“10K”，而销售额的刻度间隔是“50K”，二者相差五倍。为了更好地对比利润和销售额的实际值，需要编辑轴。

用鼠标右键单击“利润”轴，在打开的快捷菜单中选择【编辑轴】，打开“编辑轴［利润］”对话框，在“范围”选项区中选择“固定”，设置“开始”为−25 113.6286124、“结束”为 235 141.78051（注意，“利润”轴的开始值和结束值的设置与“销售额”轴的设置一致），如图 5.96 所示。或者用鼠标右键单击“利润”轴，在打开的快捷菜单中选择【同步轴】。

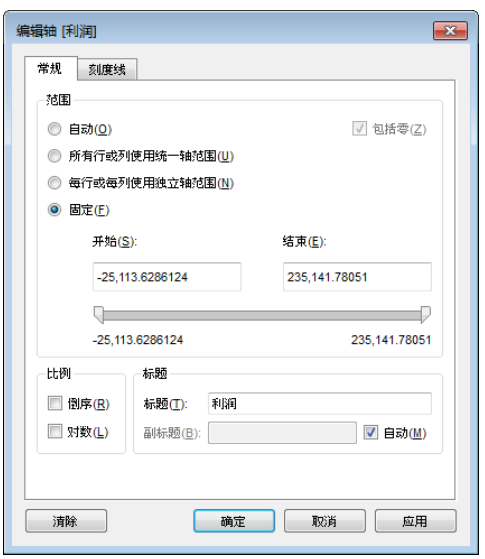


图 5.96 “编辑轴［利润］”对话框

（5）编辑图表种类。当前的双组合图是折线图和条型图，如图 5.94 和图 5.95 所示。为了美观，

可以修改图表类型，用鼠标右键单击 “利润”轴，在打开的快捷菜单中选择【标记类型】|【区域】选项，则“利润”由条型图改为面积图，如图 5.97 所示。最终效果视图按月份显示了“销售额”和“利润”的关系。

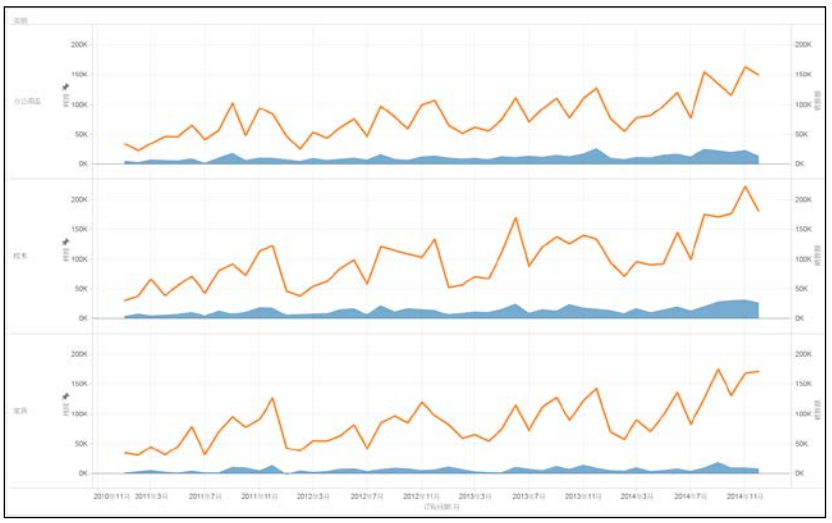


图 5.97　利润与销售额双组合图

案例 19：制作面积图“利润与销售额面积图”。

可以在“利润与销售额双组合图”的基础上编辑制作“利润与销售额面积图”。首先复制图表，然后编辑修改。

（1）复制图表。用鼠标右键单击“利润与销售额双组合图”工作表标签，在打开的快捷菜单中选择【复制工作表】选项，然后单击标签栏中的“新建工作表”图标，用鼠标右键单击新建的工作表，在打开的快捷菜单中选择【粘贴工作表】选项，操作后该工作表名称是“利润与销售额双组合图（2）”。

（2）重命名图表。用鼠标右键单击“利润与销售额双组合图（2）”工作表标签，在打开的快捷菜单中选择【重命名工作表】选项，输入新名称“利润与销售额面积图”。

（3）编辑图表类型。展开“智能显示”，单击“面积图（连续）”，效果如图 5.98 所示。再次展开“智能显示”，单击“面积图（离散）”，效果如图 5.99 所示。

图 5.98 和图 5.99 的区别在于“订购日期”是否是连续的，在甘特图、折线图和面积图中查看趋势时，“订购日期”连续十分有用。使用图 5.98 更容易发现“利润”与“销售额”的趋势。

默认情况下，将“离散”字段拖到“行”或“列”功能区时，会绘制标题，将“连续”字段添加到视图中会产生轴。字段可以在“连续”和“离散”角色之间切换。默认情况下，日期维度是离散字段，如“订购日期”字段，在将其放入功能区中时，Tableau 会自动为其选择日期级别（如年、月、周等）。

用鼠标右键单击“数据”窗格中的“订购日期”字段，在打开的快捷菜单中选择【转换为连续】选项，该字段变为绿色，再使用该字段时会自动转换为连续字段。若要恢复为离散字段，用鼠标右键单击“数据”窗格中的“订购日期”字段，在打开的快捷的菜单中选择【转换为离散】选项即可。在“数据”窗格中，蓝色字段是“离散”字段，绿色字段是“连续”字段。

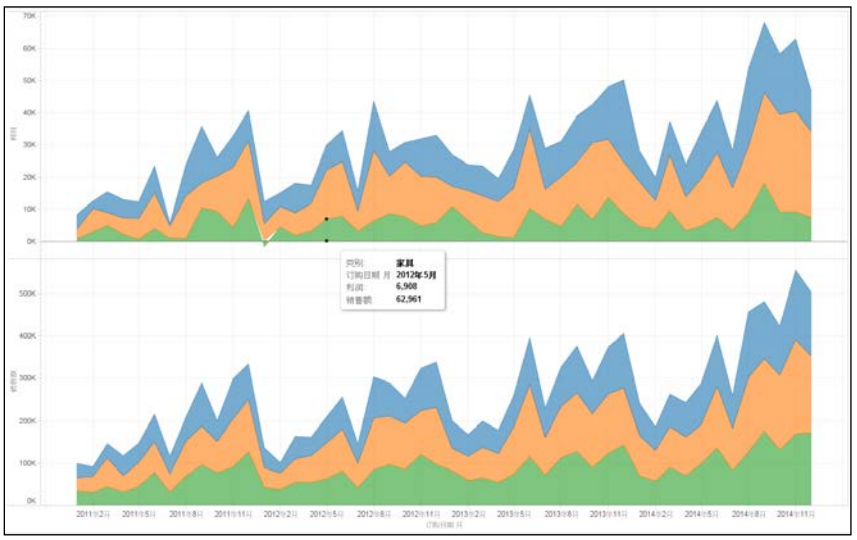


图 5.98　面积图（连续）

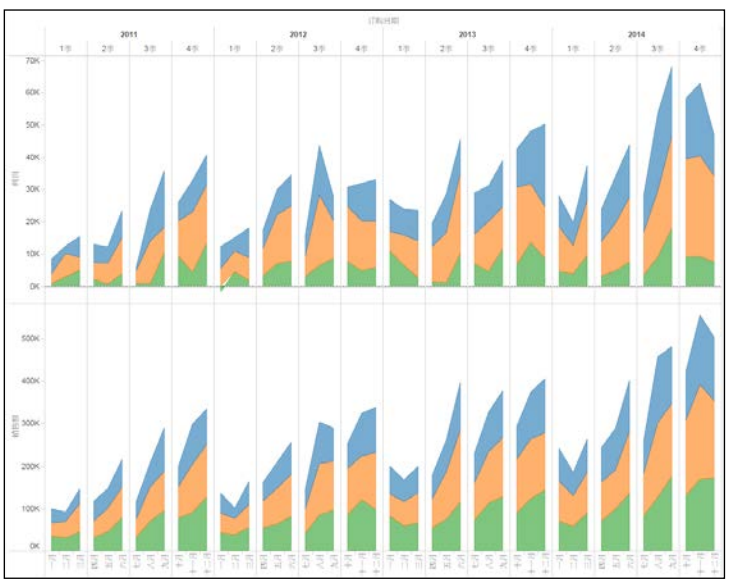


图 5.99　面积图（离散）

5.6.13　盒须图

盒须图（也称盒形图、箱图、箱线图或盒子图）可对数据的分布情况做快速而深入的分析。盒须图显示值沿轴的分布情况，盒子围住中间 50%的数据，线（也称须）可配置为在显示时包括四分位距 1.5 倍内的所有点，或配置为在数据的最大范围处显示。盒须图结构如图 5.100 所示。

上四分之一数和下四分之一数组成的矩形框是盒须图的主体，中间是数据的中位数，中位数是数据中占据中间位子的数，即数据中有一半大于中位数（在其之上），另一半小于中位数（在其之下）。上四分之一数（也称较高四分位点）表示数据中有四分之一的数大于上四分位数，即在矩形框之上；下四分之一数（也称较低四分位点）表示数据中有四分之一的数小于下四分位数，即矩形框之下。上边缘（也称上须线）是变量值本体最大值，下边缘（也称下须线）是变量值本体最小值。

案例 20：使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”，制作盒须图“地区细分市场折扣盒须图”，显示不同地区所有细分市场折扣之间的关系。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）制作盒须图。同时选择“细分市场”、“地区”维度字段和“折扣”度量字段，展开“智能显示”，选择盒须图。默认情况下，“行”功能区聚合“（总计）折扣”。

（3）解聚数据。分析度量时，如果需要在视图中独立使用度量，可以解聚数据后查看数据源的每一行数据。可以单击【分析】|【聚合度量】选项解聚数据，“行”功能区变为“折扣”。

（4）美化盒须图。将“地区”从“标记”卡拖到“列”功能区“细分市场”的右侧。用鼠标右键单击 *Y* 轴，在打开的快捷菜单中选择【编辑参考线】选项。在打开的对话框中设置“填充”为橙色、“边界”为绿色、“须状”为红色，如图 5.101 所示。

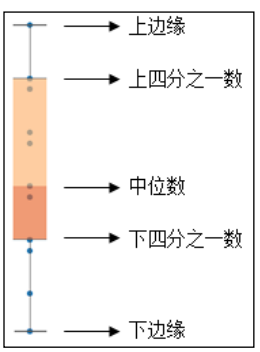


图 5.100　盒须图结构

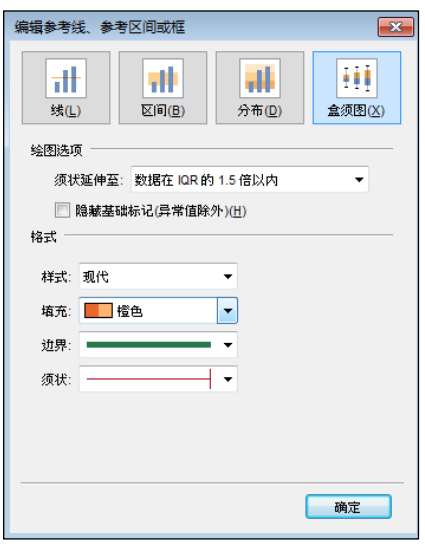


图 5.101　编辑盒须图格式

（5）设置轴格式。用鼠标右键单击 *Y* 轴，在打开的快捷菜单中选择【设置格式】选项。在打开的对话框中选择“轴”选项卡，设置数字“百分比”的小数位数是“0”。再选择“区”选项卡，单击“默认值”中“数字”右侧的三角形按钮，选择“百分比”的小数位数是“0”。

最终效果如图 5.102 所示，该图显示 EMEA（Europe、the Middle East and Africa 的字母缩写，即欧洲、中东和非洲地区）地区对于三个细分市场的折扣相同，且市场细分最大。

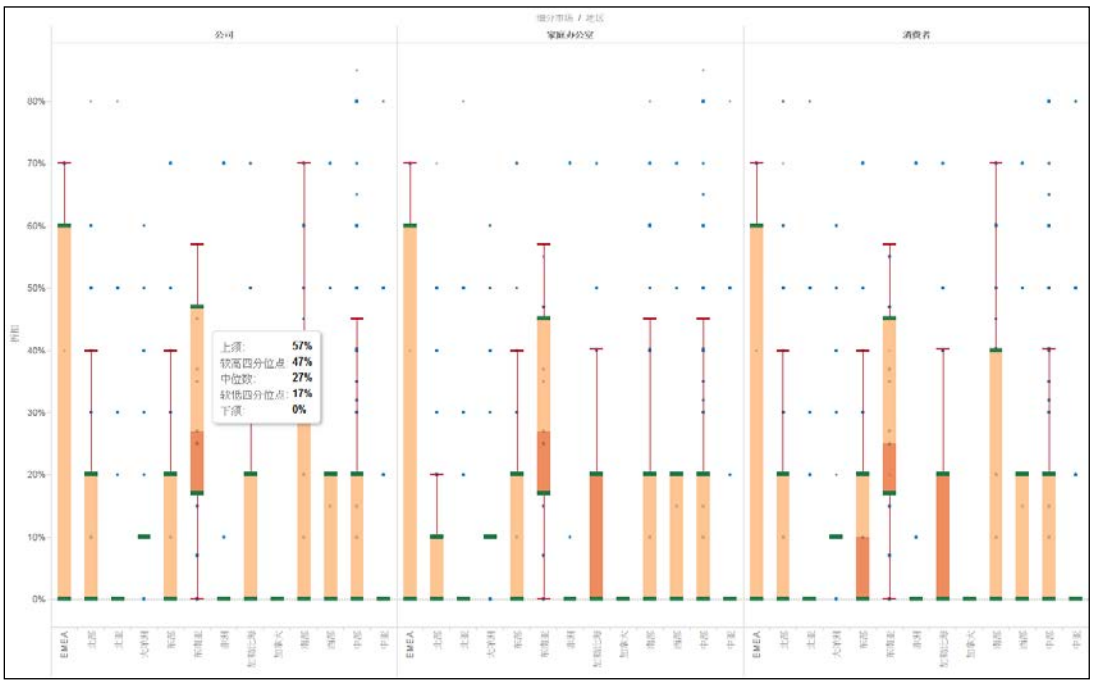


图 5.102　地区细分市场折扣盒须图

5.6.14　标靶图

标靶图是一种特殊形式的条形图，通常以定性的绩效范围（如“优”、“良”、“中”和“差”等）比较一个度量与其他度量的关系。标靶图至少包含两个度量字段。

案例 21：使用 5.6.3 小节的数据“Global Superstore_zh-cn.xlsx”和“利润计划.xlsx”，制作标靶图“利润和预计利润标靶图”，显示利润和预计利润之间的关系。

（1）连接数据。数据源连接到文件“Global Superstore_zh-cn.xlsx”。

（2）连接另一个数据源。在“已保存数据源”中打开“超市订单（Global Superstore_zh-cn）”，然后单击工具栏中的“添加新的数据源”按钮，在“已保存数据源”中打开“利润计划（利润计划）”。此时两个数据源无主次之分，并列显示在“数据”窗格中。

（3）制作标靶图。同时选择“类别”和“细分市场”维度字段并拖到“列”功能区，将“利润”和“预计利润”度量字段拖到“行”功能区，Tableau 默认将“利润”和“预计利润”聚合为“总计”，展开“智能显示”，选择标靶图。标靶图包含一条标记“预计利润”度量的平均值的参考线，还包含一个“预计利润”度量平均值 60%和 80%的参考分布。注意，快速交换两个度量的方法是用鼠标右键单击连续轴，在打开的快捷菜单中选择【交换参考线字段】选项。

（4）美化视图。单击工具栏中的“交换”按钮，实现“行”功能区和“列”功能区的字段交换。将“类别”维度字段拖到“标记”卡中的“颜色”上，将“预计利润”度量字段拖到“标记”卡中的“标签”上。

（5）编辑参考线。可以根据需求编辑参考线，用鼠标右键单击视图中的 *Y* 轴，在打开的快捷菜单中选择【编辑参考线】选项，然后选择要修改的参考线，如图 5.103 所示。然后为参考线及格式选项设置新值，如图 5.104 所示。最终效果如图 5.105 所示。

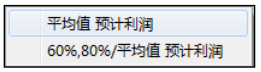


图 5.103　选择要修改的参考线

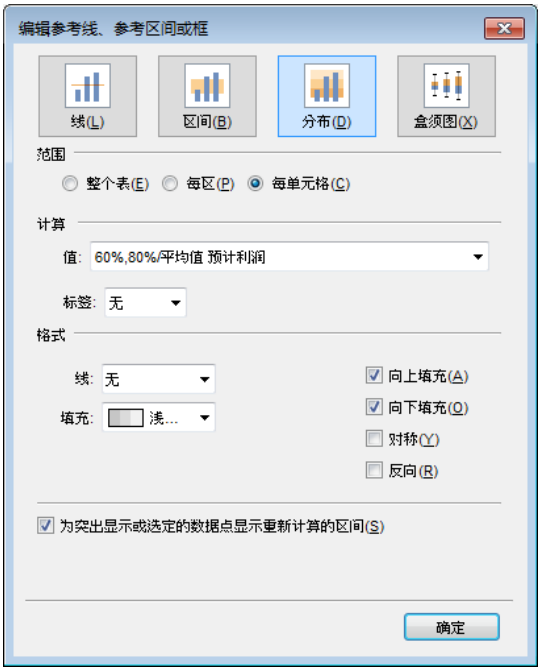


图 5.104　设置参考线及格式选项

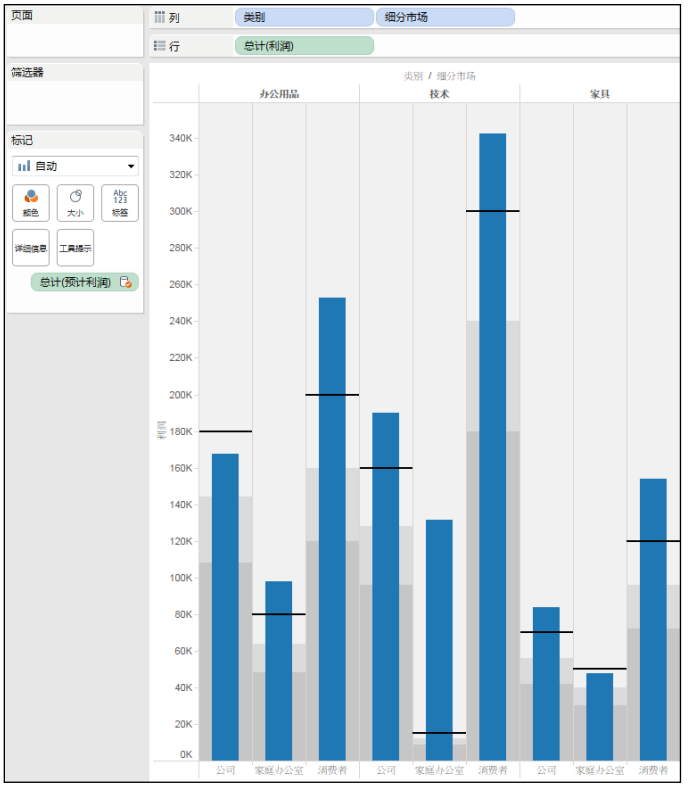


图 5.105　利润和预计利润标靶图

5.7 高级分析

高级分析包括创建自定义计算、使用内置统计数据工具、利用动态参数、分层和分组等，本节为读者介绍 Tableau 的复杂分析功能。

5.7.1　函数

Tableau 包含数字函数、字符串函数、日期函数、类型转换函数、逻辑函数、聚合函数、直通函数（RAWSQL）、用户函数和表计算函数等。其中，数字函数和字符串函数与 Excel 的相关函数类似，本小节重点介绍日期函数和聚合函数。

- 聚合函数 AVG（expression）

功能：返回表达式中所有值的平均值。AVG 只能用于数字字段，忽略 Null 值。

- 聚合函数 SUM（expression）

功能：返回表达式中所有值的总和。SUM 只能用于数字字段，忽略 Null 值。

Tableau 提供多种日期函数，具体内容如表 5.4 所示。其中，date_part 表示一个常量字符串参数。

表 5.4　日期型数据

date_part	值
'year'	四位数表示的年份
'quarter'	季节，值域 1～4
'month'	月份，值域 1～12 或“January”、“February”等
'dayofyear'	一年中的第几天，值域 1～366
'day'	天，值域 1～31
'weekday'	星期几，值域 1～7 或“Sunday”、“Monday”等
'week'	一年中的第几周，值域 1～52
'hour'	小时，值域 0～23
'minute'	分钟，值域 0～59
'second'	秒，值域 0～60

- 日期函数 DATEDIFF（date_part, date1, date2, start_of_week）

功能：返回参数 date1 和 date2 之差。

start_of_week 参数是可选参数。如果省略，一周的开始由数据源确定。例如：

DATEDIFF （'week', #2020-01-05#, #2020-01-07#, 'monday'） = 1

DATEDIFF （'week', #2020-01-05#, #2020-01-07#, 'sunday'） = 0

第一个表达式返回 1，因为当 start_of_week 为“monday”时，1 月 5 日（星期日）和 1 月 7 日（星期二）不属于同一周。第二个表达式返回 0，因为当 start_of_week 为“sunday”时，1 月 5 日（星期日）和 1 月 7 日（星期二）属于同一周。

- 日期函数 DATENAME（date_part, date, start_of_week）

功能：以字符串的形式返回 date 的 date_part。start_of_week 参数是可选参数，如果省略，一周的开始由数据源确定。

例如：DATENAME （'year', #2020-04-25#） = "2020"

DATENAME （'month', #2020-04-25#） = "April"

- 日期函数 DAY（date）

功能：以整数的形式返回参数 date 的天。

例如：DAY （#2020-04-25#） = 25

- 日期函数 MONTH（date）

功能：以整数的形式返回参数 date 的月份。

例如：MONTH （#2020-04-25#） = 4

- 日期函数 NOW（）

功能：返回系统当前日期和时间。

例如：NOW（）= 2016-09-15 3:28:25 PM

- 日期函数 TODAY（ ）

功能：返回系统当前日期。

例如：TODAY（ ）= 2016-09-15

- 日期函数 YEAR （date ）

功能：以整数的形式返回参数 date 的年份。

例如：YEAR （#2020-04-25# ） = 2020

5.7.2 聚合

Tableau 提供了一组预定义聚合，具体功能如表 5.5 所示。

表 5.5 预定义聚合

聚　合	说　明	1、2、2、3
属性	如果组中所有行都只有单个值，则返回给定表达式的值，否则显示星号 （*） 字符。忽略空值	不可使用
维度	返回度量或维度中的所有唯一值	1、2、3
总计	返回度量中数字的总和。忽略空值	8
平均值	返回度量中数字的算术平均值。忽略空值	2
中位数	返回度量中数字的中值。忽略空值	2
计数	返回度量或维度中的行数。可对数字、日期、布尔值和字符串进行计数。忽略空值	4
计数（不同）	返回度量或维度中唯一值的个数。可对数字、日期、布尔值和字符串进行计数。忽略空值	3
最小值	返回度量或连续维度中的最小数字。忽略空值	1
最大值	返回度量或连续维度中的最大数字。忽略空值	3
百分位	返回度量中指定百分位处的值。选择此聚合时，必须从提供百分位值范围的子菜单中进行选择：5、10、25、50、75、90、95。在某个字段上设置此聚合时，该字段将显示 PCT 和分配的百分比值	PCT50 的值为 2
标准偏差	基于样本总体返回给定表达式中所有值的标准差。忽略空值	0.8165
标准偏差（群体）	基于有偏差总体返回给定表达式中所有值的标准差。假定其参数由整个总体组成。此函数适用于较大的样本大小	0.7071
方差	基于样本返回给定表达式中所有值的方差。忽略空值	0.6667
方差（群体）	基于有偏差总体返回给定表达式中所有值的方差。假定其参数由整个总体组成。此函数适用于较大的样本大小	0.5000
解聚	返回基础数据源中的所有记录	1、2、2、3

5.7.3 注释

注释包含标记、点（如轴上的值或参考线）和区域（如一组分散的标记）三种。注释（也称旁注）是用来引起对视图中特定标记、点或区域的注意，通常以文本框的形式存在，并用一条线指向特定点或标记。也可以添加区域注释，作为多个标记或某个视图区域的注释。对注释可以进行编辑

内容、修改位置、定义格式和删除等操作。三种注释的含义如下。

标记注释。添加与所选标记关联的注释。只有选择标记后，此选项才可用。

点注释。为视图中特定的点添加注释。

区域注释。为视图中的区域（如视图中的一组离群点）添加注释。

案例 22: 复制 5.6.14 小节的工作表“标靶图”，并重命名为“5.7.3 注释”，为该图添加三种注释。

（1）添加标记注释。用鼠标右键单击最高利润，在打开的快捷菜单中选择【添加注释】|【标记】选项，修改“编辑注释”对话框，可以手动调整标记注释的位置、大小和指示线大小。

（2）添加点注释。用鼠标右键单击最左侧没有达到预计利润的视图，在打开的快捷菜单中选择【添加注释】|【点】选项，修改“编辑注释”对话框中的内容为“没有达到预计利润!”。可以手动调整标记点的位置、大小和指示线大小。

（3）添加区域注释。用鼠标右键单击最右侧的家具区，在打开的快捷菜单中选择【添加注释】|【区域】选项，修改“编辑注释”对话框中的内容为“家具市场整体不景气”。调整文字的大小和颜色，然后手动调整区域注释到合适位置，最终效果如图 5.106 所示。

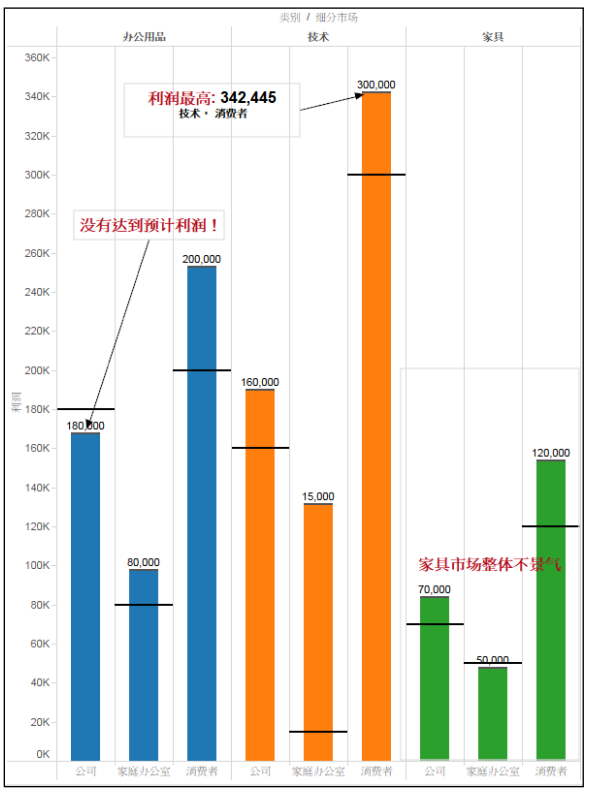


图 5.106　三种注释效果

5.7.4 计算

Tableau 计算主要分为“计算字段”和“表计算”两种。

“计算字段”是使用标准函数和运算符定义一个基于现有字段和其他计算字段公式的新字段，然后将其保存为数据的一部分。可以使用计算编辑器或通过双击功能区上的字段构建临时计算来创建计算字段。在 5.6.4、5.6.6 和 5.6.10 小节中均使用到了“计算字段”。

“表计算”是应用于整个表中值的计算，通常依赖于表结构本身。在 5.6.4 小节中使用到了“表计算”，参见图 5.57。

本小节将详细说明 Tableau 中的各种计算，主要包括计算字段的创建和编辑，临时计算的修改、添加表计算等，并使用案例说明“表计算”的八种计算类型。

1．计算字段

创建计算字段的方法主要有以下四种，可以根据实际需要选择合适的方法创建。

使用“数据”窗格创建。单击“数据”窗格上“维度”右侧的三角形按钮，在打开的下拉菜单中选择【创建计算字段】选项。

使用菜单创建。单击【分析】|【创建计算字段】选项。

使用快捷菜单创建。用鼠标右键单击“数据”窗格中的空白处，在打开的快捷菜单中选择【创建计算字段】选项。

根据字段创建。单击“数据”窗格中某个维度或度量字段，在打开的下拉菜单中选择【创建】|【计算字段】选项。

如图 5.107 所示左侧是“计算字段”对话框，右侧是提示用于完成公式的全部函数列表，并显示当前项函数的格式、功能和简单示例。

图 5.107　创建计算字段

“计算字段”对话框的顶端是创建字段的名称和新建字段保存的数据源。对话框中间是“计算编辑器”，可以输入字段（字段包含在方括号中，可以将字段从“数据”窗格或工作区的任何位置直接拖到计算编辑器，默认为橙色）、运算符（默认为黑色）、参数（具体内容参见 5.7.7 小节，默认为紫色）和注释（注释以两条正斜杠开始，直至该行结束。可编写多行注释，默认绿色）。对话框底端为提示计算是否有效，以及该新建字段影响的工作表（显示使用该字段的工作表）。

可以单击“计算字段”对话框右侧的三角形折叠按钮展开或隐藏函数列表。可以使用函数列表中的搜索框搜索函数，也可以直接输入函数。

默认“计算字段”对话框的文本较小，可以按住【Ctrl】键并使用鼠标滚动调整。但下一次打开编辑器时，文本还为默认大小。

编辑计算字段。在“数据”窗格中用鼠标右键单击要编辑的计算字段，在打开的快捷菜单中选择【编辑】选项。

2．临时计算

临时计算（也称调用类型输入计算或内联计算）是在视图功能区上可创建和更新的字段。与创建计算字段不同的是，不会为临时计算命名，但关闭工作簿时可将其保存。如果要保存临时计算以在其他工作簿的工作表中使用，可将其复制到“数据”窗格。

注意“行”、“列”、“标记”和“度量值”功能区均可以使用临时计算，但不能在“筛选器”或“页面”功能区使用。

打开 5.6.12 小节的“面积图（连续）”工作表，行工作区显示“总计（利润）”和“总计（销售额）”，双击“总计（销售额）”字段可以进行编辑，编辑后的临时计算如图 5.108 所示。也可以在功能区空白处直接双击，新建一个临时计算。

选择修改好的临时计算，按住鼠标左键拖放到“数据”窗格的维度部分，重命名后，可以新建一个计算字段。

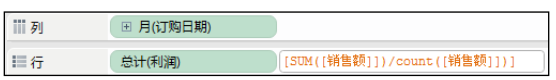


图 5.108　临时计算

3．添加表计算

用鼠标右键单击视图中的某个度量，或者单击视图中的某个度量右侧的三角形按钮，在打开的下拉菜单中选择【添加表计算】，打开“表计算［利润］”对话框，进行如图 5.109 所示的设置。单击“确定”按钮，此时度量标记为表计算（三角符号），如图 5.110 所示。

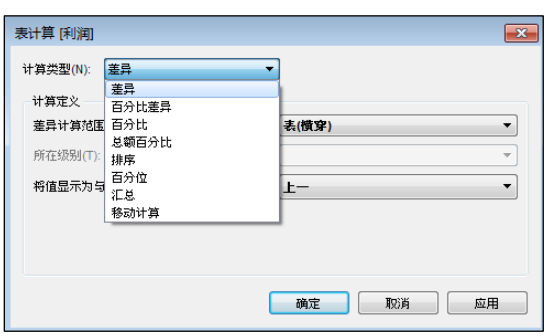


图 5.109　“表计算［利润］”对话框

图 5.110 为“总计（利润）”添加表计算

4.“表计算”的八种计算类型

下面通过案例详细说明八种计算类型的功能和使用方法。

案例 23:使用数据源文件“Global Superstore_zh-cn.xlsx”，实现八种计算。

（1）第一种计算类型为“差异”，显示绝对变化值。例如，计算每年订单利润的差异。

使用“订购日期”维度字段和“利润”度量字段制作文本表，如图 5.111 所示。

单击“标记”卡中“总计（利润）”右侧的三角形按钮，在打开的下拉菜单中选择【添加表计算】，设置“表计算[利润]”对话框中的“计算类型”为“差异”，“差异计算范围”是“订购日期”，“所在级别”是“订购日期 年”，“将值显示为与以下项的差异”是“上一”，如图 5.112 所示，单击“确定”按钮。

订购日期			
2011	2012	2013	2014
248,941	307,415	406,935	504,166

图 5.111 每年订单利润

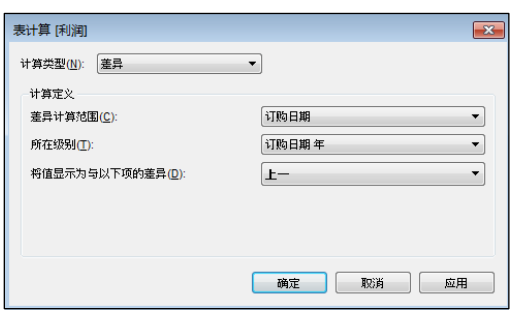


图 5.112 设置计算类型“差异”

如图 5.113 所示显示了每年与上一年订单的利润差异值。所在级别表示统计数据的级别，若选择“订购日期 季度”，则按“季度”比较与上一季度的订单的利润差异值。说明：2012 年对应的“58474”是 2012 年利润总计“307415”与 2011 年利润总计“248941”之差。

订购日期			
2011	2012	2013	2014
	58,474	99,520	97,231

图 5.113 每年订单利润“差异”

（2）第二种计算类型为“百分比差异”，用于显示变化率。若设置“表计算[利润]”对话框中的“计算类型”为“百分比差异”，“差异计算范围”是“订购日期”，“所在级别”是“订购日期 年”，“将值显示为与以下项的差异”是“下一”，单击“确定”按钮，则显示每年与下一年订单的利润百分比差异值，如图 5.114 所示。说明：2011 年对应的“–19.021%”是 2011 年利润总计与 2012 年利润总计之差“–58474”与 2012 年利润总计“307415”的百分比。

	订购日期			
	2011	2012	2013	2014
	-19.021%	-24.456%	-19.285%	

图 5.114　每年订单利润“百分比差异”

（3）第三种计算类型为“百分比”。百分比显示为与其他指定值的百分比。“百分比”与“百分比差异”的相似之处在于，可以使用它以百分比的形式计算两个值之间的变化。但是“百分比”计算绝对变化，如图 5.115 所示。说明：2012 年对应的“123.489%”是 2012 年利润总计“307415” 除以 2011 年利润总计“248941”的百分比。

	订购日期			
	2011	2012	2013	2014
		123.489%	132.373%	123.893%

图 5.115　每年订单利润“百分比”

（4）第四种计算类型为“总额百分比”，以总额百分比的形式显示值。使用“订购日期”、“类别”维度字段和“利润”度量字段制作文本表，如图 5.116 所示。

列　年(订购日期)

行　类别

	订购日期			
类别	2011	2012	2013	2014
办公用品	85,997	103,306	149,246	179,926
技术	109,247	145,977	173,627	234,928
家具	53,697	58,133	84,063	89,312

图 5.116　按类别计算每年订单利润总计

添加表计算，在打开的“表计算［利润］”对话框中设置“计算类型”为“总额百分比”，“值汇总范围”为“表（横穿）”，“寻址”设置为“计算整个表”，沿水平方向移动通过每个分区，单击“确

定”按钮，效果如图 5.117 所示。说明：2011 年办公用品对应的“16.59%”是 2011 年办公用品利润总计“85997”除以四年的办公用品利润总计“518474”的百分比。

	订购日期			
类别	2011	2012	2013	2014
办公用品	16.59%	19.92%	28.79%	34.70%
技术	16.46%	21.99%	26.16%	35.39%
家具	18.83%	20.38%	29.47%	31.32%

图 5.117　计算类型“总额百分比”表（横穿）

“值汇总范围”选项还有其他多种选择，其中“表（向下）”选项将寻址设置为对整个表计算，沿竖直方向移动通过每个分区。“表（横穿，然后向下）”选项将寻址设置为先横向后竖向计算整个表。“区（横穿）”选项是对区进行横向计算，区中横向排列的字段是寻址字段，但是分隔区的字段现在是分区字段。“区（向下）”选项将寻址设置为对表中的区向下进行计算。“区（横穿，然后向下）”选项将寻址设置为在区内横向计算，然后移至下一行继续横向进行计算。寻址字段是在表中横向排列和竖向排列的字段。“单元格”选项将寻址设置为表中的单个单元格，即所有字段都是分区字段。在计算总额百分比时，此选项最常用。

（5）第五种计算类型为“排序”，即对数值进行排序。为显示不同子类别在每年的利润排名，可以使用排序添加表计算，在打开的“表计算［利润］”对话框中，设置“计算类型”为“排序”，“计算因素”为“表（向下）”，“排序顺序”为“降序”，“将重复值评级”为“竞争排序（1，2，2，4）”，单击“确定”按钮，效果如图 5.118 所示。

列　⊞ 年(订购日期)

行　类别　子类别

		订购日期			
类别	子类别	2011	2012	2013	2014
办公用品	标签	15	15	15	15
	存储	6	7	7	7
	电器	5	6	4	4
	美术	10	10	10	10
	系固件	16	16	16	16
	信封	13	13	13	13
	用品	14	14	14	14
	纸张	9	11	9	9
	装订机	8	8	8	8
技术	电话	1	2	2	2
	复印机	2	1	1	1
	配件	7	3	6	5
	设备	11	9	11	11
家具	书架	4	5	3	3
	椅子	3	4	5	6
	用具	12	12	12	12
	桌子	17	17	17	17

图 5.118　计算类型“排序”表

重复值的排序方法共有四个选项，其功能如表 5.6 所示。

表 5.6　重复值评级选项及功能

选　项	功　能
竞争排序（1，2，2，4）	重复值的排序全部相同。重复值的下一个值的计算方式是将已经计算值的数量加 1
调整后竞争排序（1，3，3，4）	重复值的排序全部相同，计算方式是将重复值前面的值数量添加到重复值数量中。重复值的下一个值的计算方式是将已经计算值的数量加 1
密集（1，2，2，3）	重复值的排序全部相同，也就是排序序列中的下一个数字将按照重复值（就是单个值那样）计算重复值后面的下一个值
唯一（1，2，3，4）	将根据计算排序的方向为重复值指定不同的排序

（6）第六种计算类型为“百分位”，用于计算百分位值。为了查看不同年份各种子类别的利润是否有波动，可以使用“百分位”计算类型添加表计算，在打开的“表计算［利润］”对话框中设置“计算类型”为“百分位”，“计算因素”为“表（向下）”，“排序顺序”为“升序”，单击“确定”按钮，效果如图 5.119 所示。可以看出，2011 ~ 2014 年的 4 年中，各种子类别的利润整体比较稳定。

		订购日期			
类别	子类别	2011	2012	2013	2014
办公用品	标签	17.65%	17.65%	17.65%	17.65%
	存储	70.59%	64.71%	64.71%	64.71%
	电器	76.47%	70.59%	82.35%	82.35%
	美术	47.06%	47.06%	47.06%	47.06%
	系固件	11.76%	11.76%	11.76%	11.76%
	信封	29.41%	29.41%	29.41%	29.41%
	用品	23.53%	23.53%	23.53%	23.53%
	纸张	52.94%	41.18%	52.94%	52.94%
	装订机	58.82%	58.82%	58.82%	58.82%
技术	电话	100.00%	94.12%	94.12%	94.12%
	复印机	94.12%	100.00%	100.00%	100.00%
	配件	64.71%	88.24%	70.59%	76.47%
	设备	41.18%	52.94%	41.18%	41.18%
家具	书架	82.35%	76.47%	88.24%	88.24%
	椅子	88.24%	82.35%	76.47%	70.59%
	用具	35.29%	35.29%	35.29%	35.29%
	桌子	5.88%	5.88%	5.88%	5.88%

图 5.119　计算类型“百分位”表

（7）第七种计算类型为“汇总”，用于显示累积总额，可沿维度或表结构计算累计总额。下面计算每年利润总和并添加表计算，在打开的“表计算［利润］”对话框中设置“计算类型”为“汇总”，“根据以下因素汇总值”为“总计”，“计算因素”为“表（横穿）”，单击“确定”按钮，效果如图 5.120 所示。2014 年的“518474”是 2011～2014 年利润总计之和（参考图 5.116 中的数据）。

	订购日期			
类别	2011	2012	2013	2014
办公用品	85,997	189,302	338,548	518,474
技术	109,247	255,224	428,851	663,779
家具	53,697	111,830	195,893	285,205

图 5.120　计算类型“汇总”表

（8）第八种计算类型为“移动计算”。通常用于平滑短期数据波动，可以查看长期趋势，如查看证券数据、市场行情等。如图 5.121 所示显示了 2014 年全年按订购日期总计利润的情况，用户很难在视图中分析出趋势。添加“移动计算”后就会方便对视图的理解。

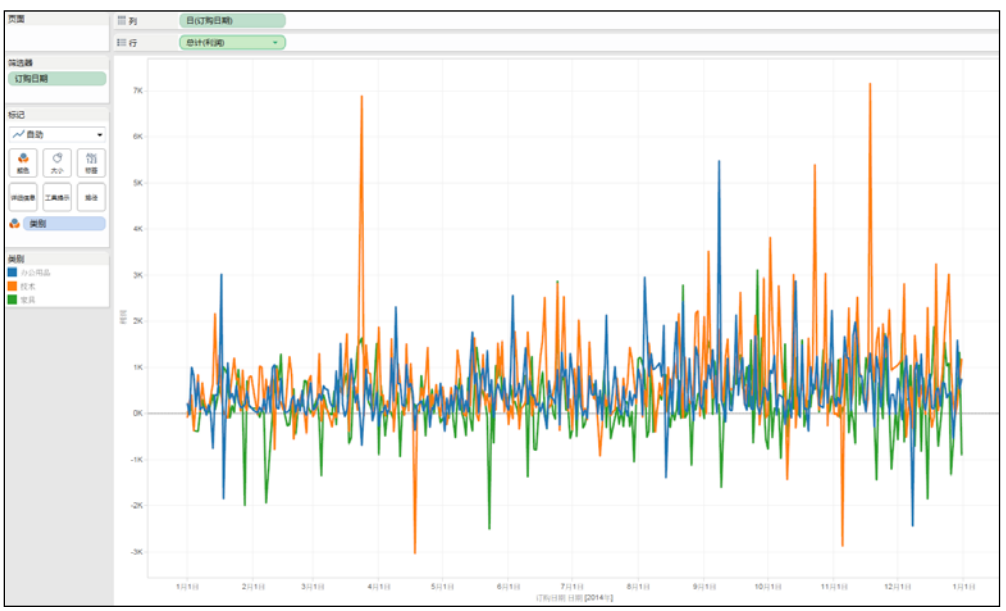


图 5.121　2014 年全年按订购日期总计利润

为“行”功能区的“总计（利润）”添加表计算，在打开的“表计算［利润］”对话框中设置“计算类型”为“移动计算”，“根据以下因素汇总值”为“平均值”，“移动计算因素”为“表（横穿）”，“前面的值”为“3”，“以后的值”为“4”，勾选“包括当前值”复选框，如图 5.122 所示，单击“确定”按钮。

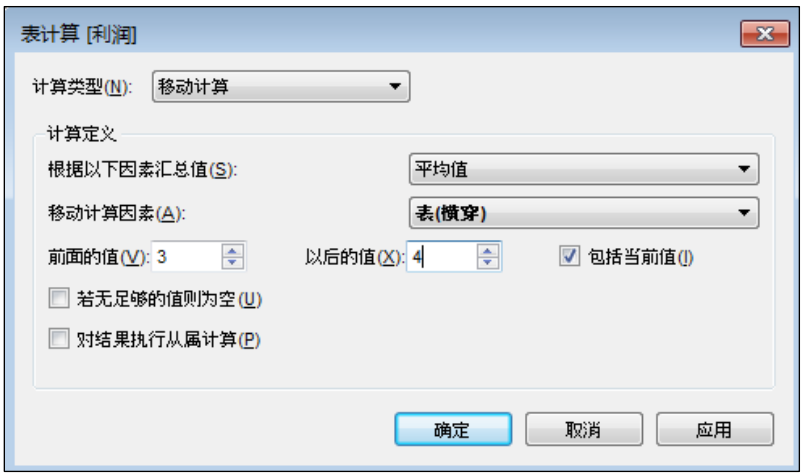


图 5.122　设置“表计算［利润］”对话框

如图 5.123 所示为沿着视图中的行将这些值汇总为平均值。每个值都是围绕当前值的七天（之前三天，之后四天）的平均值。从该图可以明显地看出三种类别的利润趋势，技术类产品利润增加的趋势最为明显。

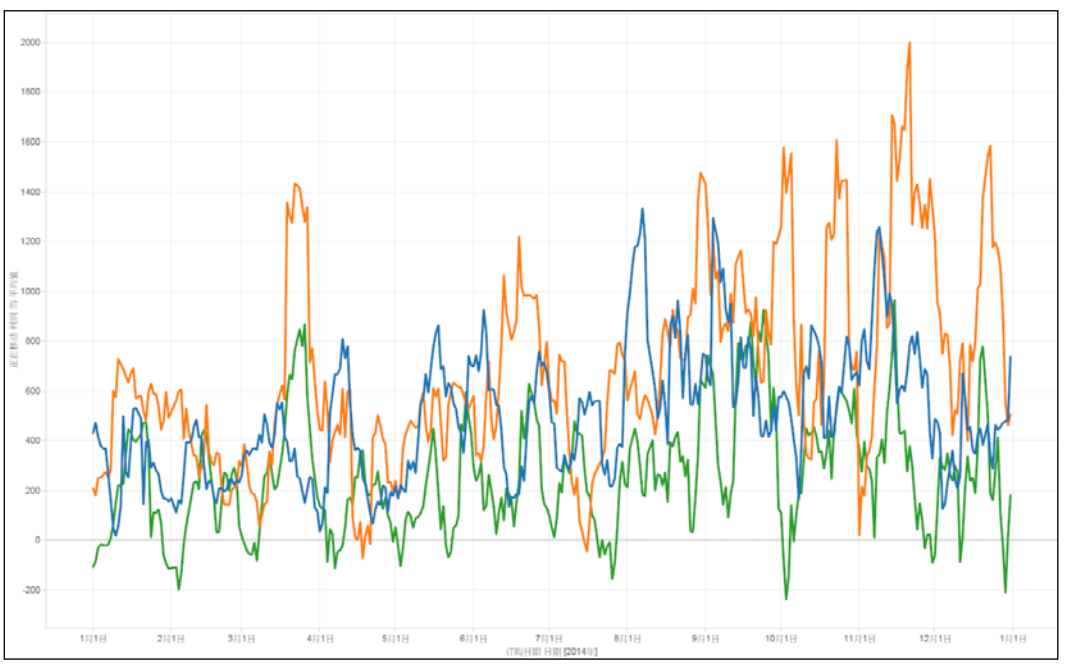


图 5.123　计算类型“移动计算”表

5.7.5 简单预测

预测是根据已知数据和当前趋势对未来趋势进行的计算。Tableau 中使用预测时要求视图中要至少包含一个日期字段和一个度量字段。已有 2011 ~ 2014 年（按月）订购日期总计的利润视图，单击【分析】|【预测】|【显示预测】选项，效果如图 5.124 所示。

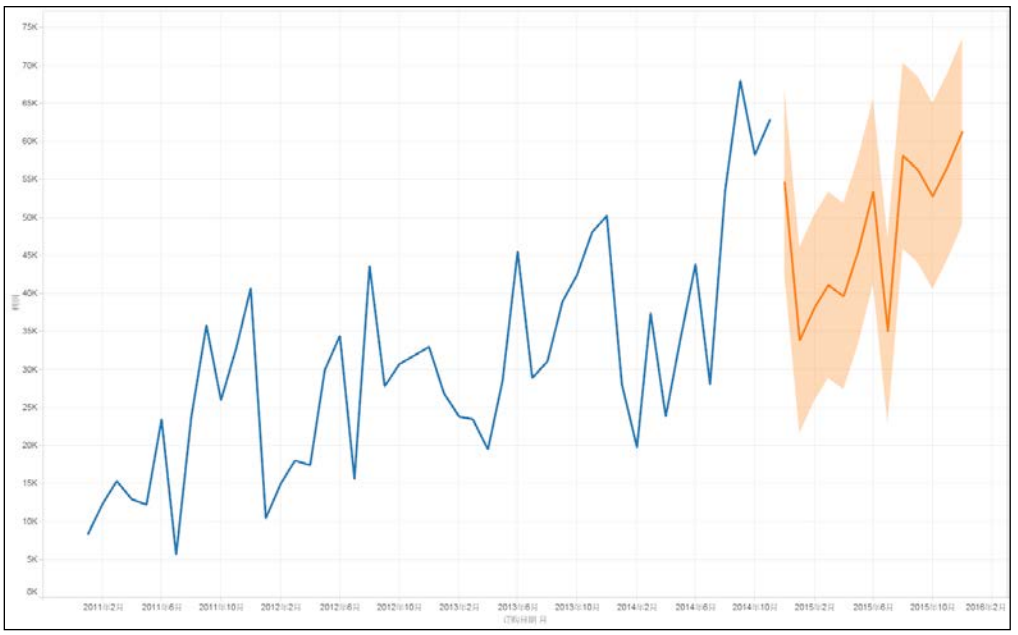


图 5.124　简单预测效果

如图 5.124 所示显示了 2015 年的预测利润，预测值比原始数据的颜色更浅。

单击【分析】|【预测】|【预测选项】选项可以修改预测参数。单击【分析】|【预测】|【描述预测】选项可以查看预测模型，了解用于创建预测的参数。

5.7.6 合计

合计包含总计和小计两种。总计对“行”或“列”计算总和。小计对选定的维度或所有符合条件的维度计算总和。

计算利润总计。已有按“类别”和“邮寄方式”统计的利润总和文本表视图，单击【分析】|【合计】|【显示行总计】选项可以启用行总计，单击【分析】|【合计】|【显示列总计】选项可以启用列总计，效果如图 5.125 所示。

计算总计时必须保证视图中“列”功能区或“行”功能区中有一个维度。如果显示列标题，则计算列总计。如果显示行标题，则计算行总计。如图 5.125 所示既有行标题，也有列标题，所以行列总计均有。总计不能应用于连续维度。

首次使用总计时，将对解聚数据计算总计。如计算利润均值总计时，是对解聚后的所有符合条件的数据求均值。

计算利润均值总计。启用行列总计，修改“标记”卡中的“文本”为“平均值（利润）”，并设置格式为“货币（标准）”，效果如图 5.126 所示。

	类别			
邮寄方式	办公用品	技术	家具	总计
标准级	307,252	405,461	177,883	890,596
当日	28,335	29,041	18,797	76,173
二级	111,872	137,995	42,717	292,584
一级	71,015	91,282	45,808	208,105
总计	518,474	663,779	285,205	1,467,457

图 5.125　行、列利润总计

	类别			
邮寄方式	办公用品	技术	家具	总计
标准级	¥16.44	¥66.28	¥29.80	¥28.94
当日	¥16.92	¥56.06	¥37.00	¥28.20
二级	¥17.73	¥67.98	¥21.71	¥28.38
一级	¥15.44	¥61.84	¥32.01	¥27.73
总计	¥16.58	¥65.45	¥28.88	¥28.61

图 5.126　行、列利润均值总计（解聚数据）

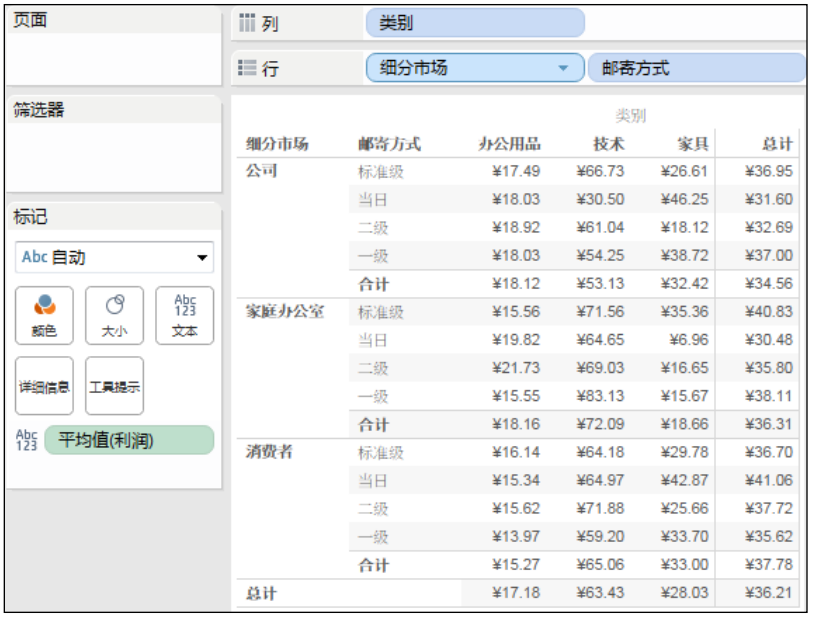
视图的第一行包含 4 个数据，分别是“¥16.44”、“¥66.28”、“¥29.80”和“¥28.94”，但“¥16.44”、“¥66.28”和“¥29.80”的均值并不是“¥37.51”，而是“¥28.94”。原因是“总计”的计算方法是解聚数据，即统计邮寄方式是“标准级”的所有原始数据总和是“890596”，然后再计数是“30775”，最后求出均值是“¥28.94”。

若希望得到“¥16.44”、“¥66.28”和“¥29.80”的均值，可以在图 5.126 的基础上，单击【分析】|【合计】|【全部汇总依据】|【平均值】选项，效果如图 5.127 所示。这种汇总称为双步汇总，因为在总计列中看到的平均值被聚合了两次，第一次聚合获得了列值或行值，第二次聚合跨列或跨行得出了总计。

	类别			
邮寄方式	办公用品	技术	家具	总计
标准级	¥16.44	¥66.28	¥29.80	¥37.51
当日	¥16.92	¥56.06	¥37.00	¥36.66
二级	¥17.73	¥67.98	¥21.71	¥35.80
一级	¥15.44	¥61.84	¥32.01	¥36.43
总计	¥16.63	¥63.04	¥30.13	¥36.60

图 5.127　行、列利润均值总计（非解聚数据）

计算小计。任何数据视图均可以包括小计。经常使用菜单项为所有字段添加小计，单击【分析】|【合计】|【添加所有小计】选项，效果如图 5.128 所示。



		类别			
细分市场	邮寄方式	办公用品	技术	家具	总计
公司	标准级	¥17.49	¥66.73	¥26.61	¥36.95
	当日	¥18.03	¥30.50	¥46.25	¥31.60
	二级	¥18.92	¥61.04	¥18.12	¥32.69
	一级	¥18.03	¥54.25	¥38.72	¥37.00
	合计	¥18.12	¥53.13	¥32.42	¥34.56
家庭办公室	标准级	¥15.56	¥71.56	¥35.36	¥40.83
	当日	¥19.82	¥64.65	¥6.96	¥30.48
	二级	¥21.73	¥69.03	¥16.65	¥35.80
	一级	¥15.55	¥83.13	¥15.67	¥38.11
	合计	¥18.16	¥72.09	¥18.66	¥36.31
消费者	标准级	¥16.14	¥64.18	¥29.78	¥36.70
	当日	¥15.34	¥64.97	¥42.87	¥41.06
	二级	¥15.62	¥71.88	¥25.66	¥37.72
	一级	¥13.97	¥59.20	¥33.70	¥35.62
	合计	¥15.27	¥65.06	¥33.00	¥37.78
总计		¥17.18	¥63.43	¥28.03	¥36.21

图 5.128　利润均值总计和小计

若维度较多，可以仅为选定的维度进行小计。首先确保选定的维度已经启用了“总计”，然后在“行”功能区或“列”功能区中用鼠标右键单击选定的维度，在打开的快捷菜单中选择【小计】选项即可对选定的维度启用小计。

5.7.7　参数

参数是可在计算中替换常量值的一个变量。如根据数据“省会城市空气质量.xlsx”制作的视图，可以依据 AQI 值确定某天哪些省会城市适合人们进行户外活动，此处的 AQI 值可以设置为一个常量“50”，即 AQI 值低于“50”的城市适合人们进行户外活动，但很多时候可以将这个常量用一个可变的量替代，并显示参数控件，用户可以设置该变量。

制作符号地图“5.7.7 参数”。选择“省会城市”维度字段和“AQI 指数”度量字段，展开“智能显示”，选择“符号地图”。将“日期”维度字段拖到“筛选器”选项卡中并设置为“天”，单击“下一步”按钮并选择“1”，即仅筛选“1 日”的信息。用鼠标右键单击“筛选器”选项卡中的“(日)日期”，在打开的快捷菜单中选择“显示快速筛选器”。单击“(日)日期”右上角的三角形按钮，在打开的下拉菜单中选择“单值(滑块)”。

创建计算字段“适合户外活动”。单击“数据”窗格中的“AQI 指数”度量字段，在打开的下拉菜单中选择【创建】|【计算字段】选项。创建的字段名称是“适合户外活动”，在“计算编辑器”中输入“IF ([AQI 指数]) <= 70 THEN"Yes"ELSE"No"END”，如图 5.129 所示。

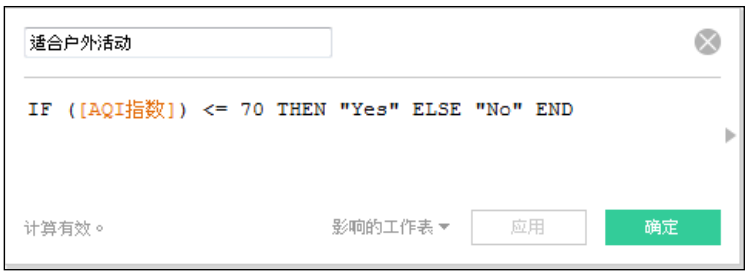


图 5.129　新建计算字段“适合户外活动”

注意：虽然字段名或字符串均可以是中文，但依然建议“计算编辑器”中除方括号中的字段名外，其他所有内容均为英文半角，特别注意引号、四则运算符号、函数的圆括号等是英文的。

创建计算参数的方法主要有以下三种，可以根据实际需要选择合适的方法创建。

- 使用“数据”窗格创建。单击“数据”窗格上“维度”右侧的下拉按钮，在打开的下拉菜单中选择【创建参数】选项。
- 根据字段创建。单击“数据”窗格中某个维度或度量字段，在打开的下拉菜单中选择【创建】|【参数】选项。
- 使用“筛选器”创建。在“筛选器”对话框的“顶部”选项卡中，“按字段”或“按公式”选择“创建新参数”，如图 5.130 所示。

图 5.130　使用“筛选器”创建参数

在符号地图“5.7.7 参数”中使用“数据”窗格创建参数，打开“创建参数”对话框，设置“名称”为“设置 AQI 指数”，填写注释信息描述该参数；“数据类型”设置为“整数”，“当前值”为描述参数的初始值，设置为“0”；“显示格式”用来设置值的格式，设置为“自动”；“允许的值”可以定义为“全部”、“列表”或“范围”，本案例中选择“范围”；设置最小值为“0”、最大值为“200”、步长为“5”，如图 5.131 所示。

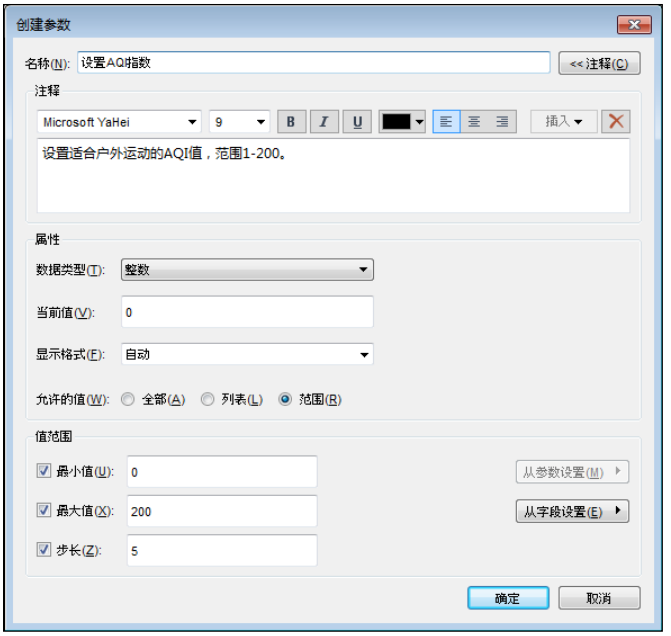


图 5.131 “创建参数”对话框

在“创建参数”对话框中，“允许的值”有三个选项，各选项的含义如下。

“全部”：表示参数控件是字段中的简单类型。

“列表”：表示参数控件提供一个可供选择的值列表。

“范围”：表示参数控件可用于选择指定范围中的值。

使用参数。将新建的参数“设置 AQI 指数”连接到计算字段“适合户外活动”。在“数据”窗格中用鼠标右键单击“适合户外活动”维度字段，在打开的快捷菜单中选择【编辑】选项，在打开的“计算字段”对话框中将常量“70”修改为“IF （[AQI 指数]） <= [设置 AQI 指数] THEN"Yes"ELSE"No"END”，如图 5.132 所示。

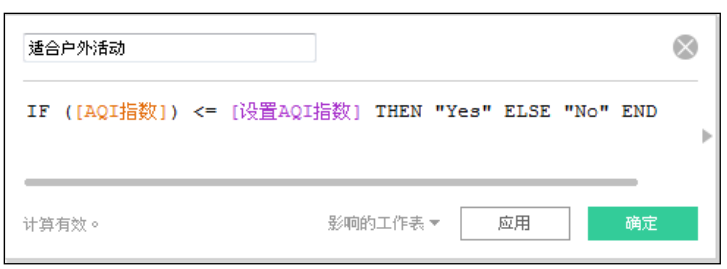


图 5.132 用参数替换计算字段“适合户外活动”中的常量

显示参数控件。在“数据”窗格中用鼠标右键单击参数“设置 AQI 指数”，在打开的快捷菜单中选择【显示参数控件】选项。

默认情况下，参数控件显示在右侧。用鼠标拖动参数控件和快速筛选器到左下角，美化视图，最终视图效果如图 5.133 所示。手动调整日期，设置 AQI 指数参数后，AQI 值小于该参数的城市是橙色的，反之是蓝色的。城市的颜色与选择的日期和设置的 AQI 指数相关。

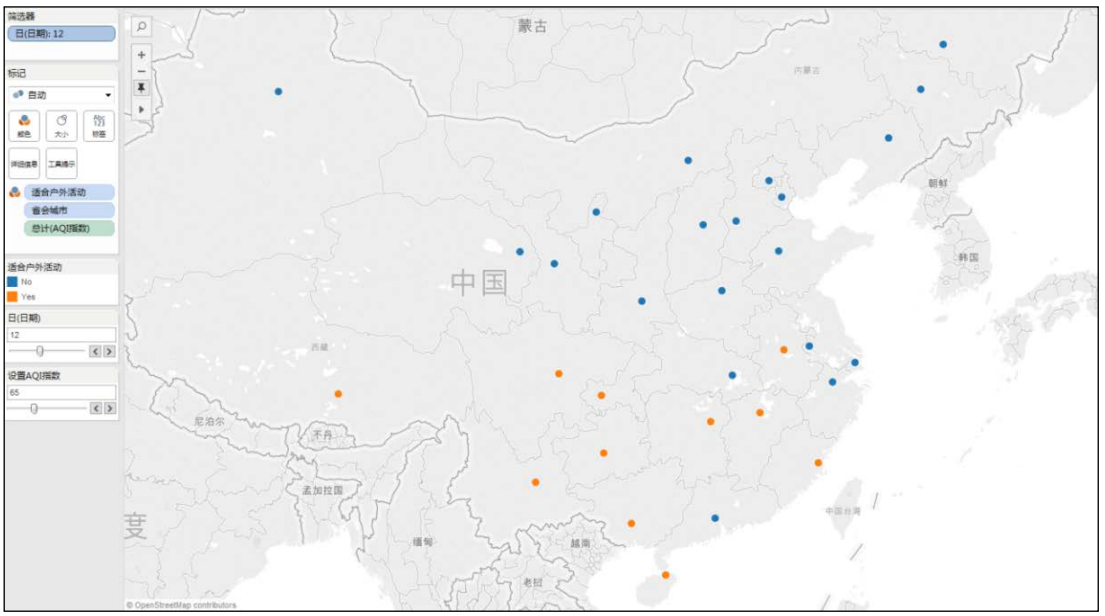


图 5.133　使用参数的符号地图

参数的使用方法与常量值类似，既可以在计算中使用参数，也可以在筛选器、参考线中使用参数。

5.7.8　分层

连接数据“Global Superstore_zh-cn.xlsx”。新建工作表，将“订购日期”拖到“行”功能区，Tableau 会自动将该字段分隔为分层结构，即默认情况下该字段以“年（订购日期）”的方式呈现数据，单击字段前面的加号（ + ）可以按季度、月和日等方式细分视图，这就是典型的分层结构。

用户可以根据需要创建自定义的分层结构，创建分层的方法有以下两种。

使用“数据”窗格创建。在“数据”窗格中选择“国家/地区”、“州”和“城市”三个字段，单击鼠标右键，在打开的快捷菜单中选择【分层结构】|【创建分层结构】选项，打开“创建分层结构”对话框，如图 5.134 所示。

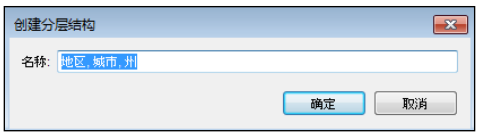


图 5.134 “创建分层结构”对话框

手动创建。在“数据”窗格中，选择一个字段，按住鼠标左键拖动到“数据”窗格中其他字段的上方即可创建分层结构。可以通过拖放对分层结构的各层重新排序。

无论使用哪种方法创建分层，创建后的效果均相同，如图 5.135 所示。可以将其他字段拖到已创建好的分层结构中，也可以在分层结构中拖动字段重新排序，如图 5.136 所示。

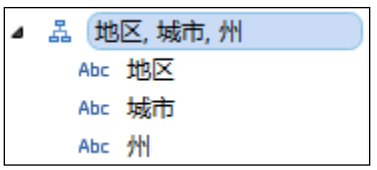


图 5.135 分层结构图

图 5.136 调整顺序后的分层结构图

拖动创建好的分层到“行”或“列”功能区，可以通过单击“+”或“–”按钮进行“下钻”或“上钻”。“下钻”或“上钻”均对数据进行了筛选，视图呈现的数据是不同的，如图 5.137 所示。

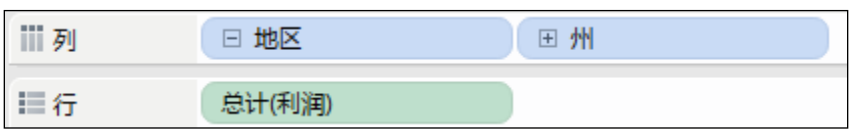


图 5.137 下钻或上钻

5.7.9 分组

分组是将维度的某几个数据成员合并为一个类别，如将“子类别”中的“标签”、“信封”和“纸张”合并为“办公小件”分组。

创建分组。选择要分组的一个或多个成员。单击工具栏中的“组成员”图标，默认分组名称是“子类别”，因为选择的三个成员来源于“子类别”，如图 5.138 所示。

还可以在视图中选择内容，然后单击鼠标右键，在打开的快捷菜单中选择【组】选项来创建分组或者从工具提示上选择【组成员】。

重命名分组。在“数据”窗格中用鼠标右键单击分组字段“子类别”，在打开的快捷菜单中选择【重命名】或【编辑组】选项，将字段名称修改为“办公小件”。

编辑分组。在“编辑组［办公小件］”对话框中，如图 5.139 所示，将列表框中的选定成员拖放到组中即可添加成员，从列表框中拖出选定成员即可删除成员。

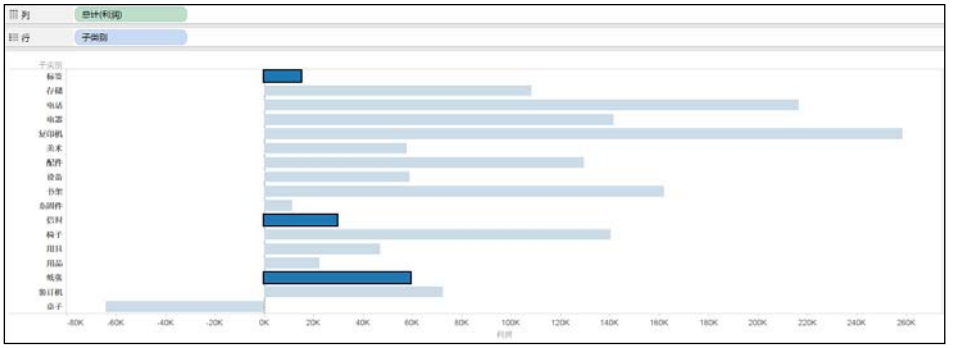


图 5.138　创建分组

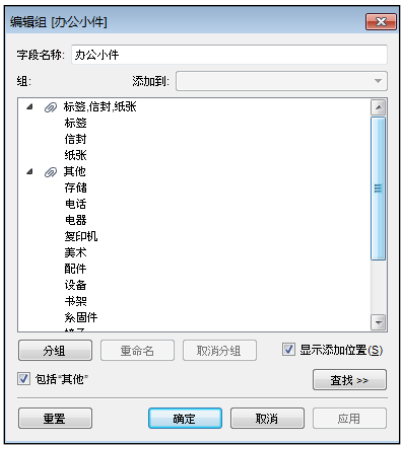


图 5.139　编辑分组

所选标记将分成一组，而所有其他成员将合并到“其他”类别中，新的组字段将自动添加到“标记”卡的“颜色”上。单击“查找”按钮，可以在展开的“查找成员”文本框中搜索和选择符合条件的成员，这对包含大量成员的维度是非常有用的。

5.7.10　“页面”功能区

默认情况下，“页面”功能区在视图的左上角，可以将视图划分为一组多个页面，每个页面包含不同的视图，方便用户更准确地分析特定字段对视图中其他字段的影响。

将某个维度字段放到“页面”功能区时，将为该维度字段的每个成员添加一个新行。将某个度量字段放到“页面”功能区时，该度量将会自动转换为离散度量。使用将字段移到“页面”功能区时添加到视图中的控件，用户可以在一个公共轴上翻阅每页视图并进行比较。

新建一个视图，“列”功能区包含“总计（利润）”字段，“行”功能区包含“日（订购日期）”字段和“总计（销售额）”字段，用“类别”字段区分颜色，如图 5.140 所示。

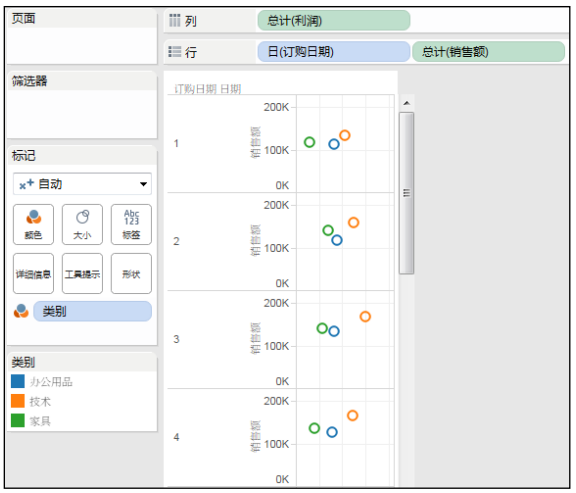


图 5.140　原始无“页面”功能区

为查看每天的信息，用户需要使用右侧的滚动条。可以将“行”功能区的“日（订购日期）”字段拖到“页面”功能区，这时“页面”功能区下方自动添加了页面控件。可以使用手动翻阅页面、跳转到特定页面或者自动翻阅页面三种方式，在一组页面之间浏览，如图 5.141 所示。

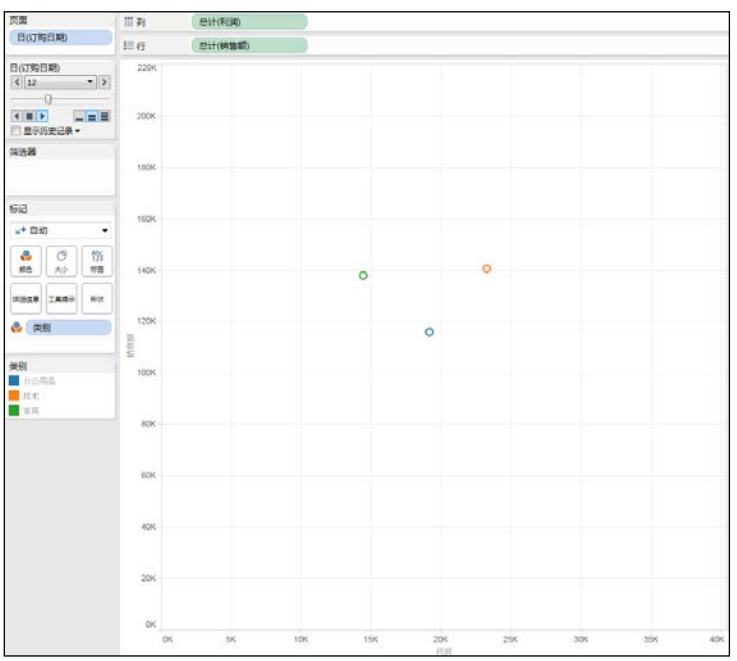


图 5.141　增加“页面”功能区

5.7.11　数据桶和直方图

数据桶是将原始数据中的某个度量字段的值分成不同的类（类似于数据挖掘中的分类），如按照某种方式将销售额分类。在使用数据创建直方图时，默认情况下，Tableau 自动创建数据桶，也可以手动创建分桶制作直方图。只有关系型数据源中的数据可以分桶，而且只有度量字段可以分桶。

创建数据桶。在“数据”窗格中，选择“装运成本”度量字段，用鼠标右键单击该字段或者单击该字段右侧的三角形按钮，在打开的下拉菜单中选择【创建】|【数据桶】选项，打开“创建级[装运成本]”对话框，设置分桶的属性，如图 5.142 所示。

图 5.142　“创建级[装运成本]”对话框

在“创建级[装运成本]”对话框中输入新字段的名称。为帮助确定最佳数据桶大小，单击“加载”按钮显示该度量字段的系列值。单击“确定”按钮，创建的数据桶被显示在“数据”窗格的“维度”区域，因为数据桶字段是离散的字段。

将“装运成本”度量字段拖到“行”功能区，将“装运成本（数据桶）”字段拖到“列”功能区，创建的直方图如图 5.143 所示，默认情况下分为 10 桶。

也可以直接指定或修改数据桶的大小。用鼠标右键单击“装运成本（数据桶）”字段或单击该字段右侧的三角形按钮，在下拉菜单中选择【编辑】选项，在打开的“编辑级[装运成本]”对话框中输入数据桶大小为“15”。直方图效果如图 5.144 所示，增加了桶数。

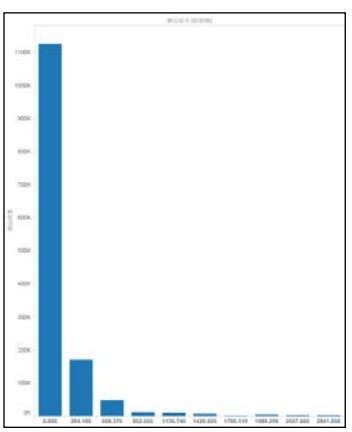


图 5.143　默认数据桶大小

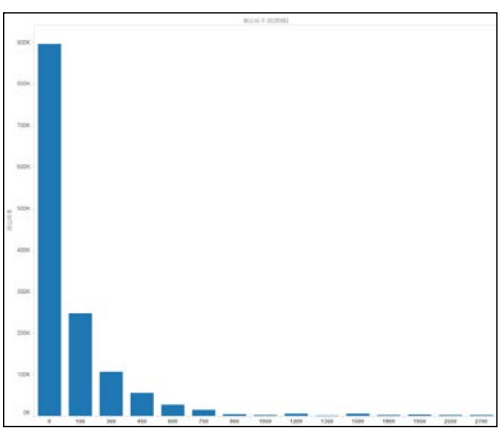


图 5.144　修改数据桶大小

默认情况下所有数据桶的大小是相等的。每个数据桶相当于一个容量相同的容器，汇总特定数据。数据桶的标签标明该数据桶所包含数字范围的下限。如图 5.144 所示数据桶大小是“150”，则标签为“300”的数据桶包含的数据值范围是大于或等于 300 但小于 450。

5.7.12 背景图像

背景图像是视图中显示在数据下面的图片。Tableau 允许用户使用在线或离线提供商提供的动态地图，也允许用户使用背景图像设置个性化地图，如使用 NBA 赛场图片、建筑物分布图、海底模型、用于分析 Web 日志的网页图像或仓库图等。

案例 24： 制作仓库背景图像。

（1）制作背景图像。打开 Excel，使用单元格合并和颜色填充等功能制作一个简单的背景图像，如图 5.145 所示。使用画图、Photoshop 等图像编辑软件，将其保存为图像“库房图.png”。注意，为了图片美观不需要保留图 5.145 中的列头和行头信息。

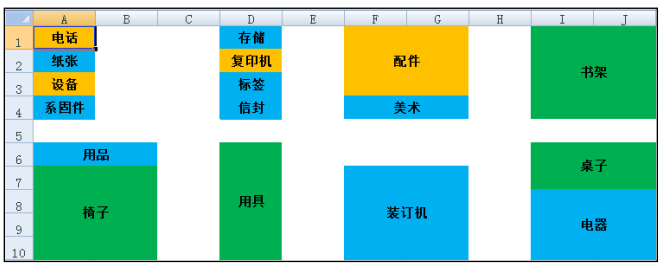


图 5.145 用 Excel 制作背景图像

（2）设置 *X* 轴和 *Y* 轴。根据物品在库房中保存的位置，设置相应的 *X* 轴和 *Y* 轴的值。本案例中假设图片行范围和列范围均是 0～10，即左下角是（0,0），右上角是（10,10），所以“电话”的中心位置是（0.5,9.5），所有子类别 *X* 轴和 *Y* 轴的值如图 5.146 所示。

	A	B	C
1	子类别	X	Y
2	电话	0.5	9.5
3	纸张	0.5	8.5
4	设备	0.5	7.5
5	系固件	0.5	6.5
6	用品	1	4.5
7	椅子	1	2
8	存储	3.5	9.5
9	复印机	3.5	8.5
10	标签	3.5	7.5
11	信封	3.5	6.5
12	用具	3.5	2.5
13	配件	6	8.5
14	美术	6	6.5
15	装订机	6	2
16	书架	9	8
17	桌子	9	4
18	电器	9	1

图 5.146 子类别 *X* 轴和 *Y* 轴的值

（3）保存文件。将保存所有子类别 *X* 轴和 *Y* 轴值的工作表复制到数据文件“Global Superstore_zh-cn.xlsx”，并另存为“Global Superstore_zh-cn（包含 *XY* 轴）.xlsx”。

（4）连接数据。数据源连接到文件“Global Superstore_zh-cn（包含 *XY* 轴）.xlsx”。如图 5.147 所示。

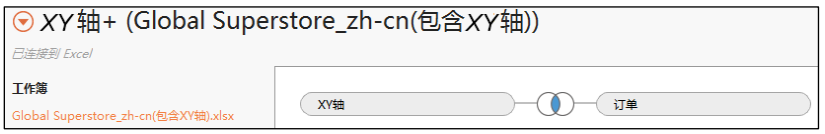


图 5.147　连接数据

（5）添加背景图像。单击【地图】|【背景图像】选项，在打开的“背景图片”对话框中单击“添加图像”按钮，在打开的“添加背景图片”对话框中，输入名称“库房图”，单击“浏览”按钮后在弹出的对话框中选择要添加到背景中的图像“库房图.png”，然后选择要映射到图像 *X* 轴的字段“X”，设置左值是“0”、右值是“10”，再选择 *Y* 轴的字段“Y”，设置下值是“0”、上值是“10”，如图 5.148 所示，单击“确定”按钮。

注意，“添加背景图片”对话框中的“冲蚀”滑块用于调整图片的浓度，滑块越向右移，图像越淡。在该对话框的“选项”选项卡中可以勾选“锁定纵横比”，保证图像的原始长宽比例。

图 5.148　“添加背景图片”对话框

（6）在视图中显示背景图像。新建工作表“库房图”，将“X”和“Y”度量字段分别拖到“列”功能区和“行”功能区，单击【分析】|【聚合度量】选项解聚所有度量。将“子类别”拖到“标记”卡的“详细信息”上（注意：这是非常关键的一步，因为在 5.8.5 小节中要添加“子类别”筛选操作）。

取消背景图像中的小圆圈。完成上述步骤后，每个“子类别”名称的中心均有一个小圆圈。单

击“标记”卡中的“颜色”，设置透明度为“0%”，最终效果如图 5.149 所示。

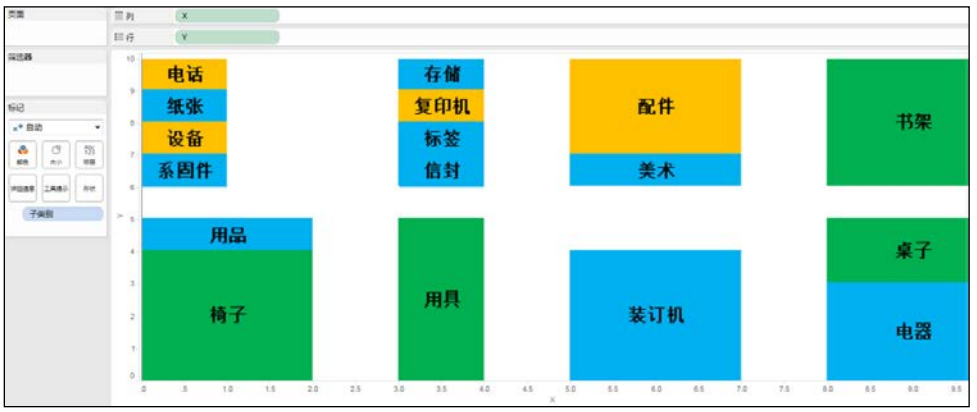


图 5.149　添加背景图像的最终效果

5.8 仪表板

仪表板包含多个视图、对象、图例或快速筛选器等，其中视图是仪表板最重要的组成部分。一般情况下，仪表板中包含一个或多个相关的工作表视图，可以方便地汇总、对比和浏览数据表。创建仪表板时，可以选择工作簿中的任何一个或多个工作表添加到视图，也可以为仪表板添加文本区域、网页和图像等对象。仪表板中的视图连接至它们表示的工作表，所以仪表板的内容会在更改工作表后实时更新，对仪表板进行的多项更改也会影响其包含的工作表。

从仪表板可以方便地转到选定工作表的原始编辑视图，也可以在仪表板中复制工作表或隐藏仪表板。

5.8.1　创建仪表板

仪表板的创建包含新建仪表板、向仪表板中添加工作表视图和添加其他仪表板对象（如图像、文本或网页等）。

新建仪表板。单击【仪表板】|【新建仪表板】选项，或者单击标签栏中的“新建仪表板”图标，工作表的底部标签栏处会显示新建的仪表板标签，默认名称为“仪表板 1”。

仪表板工作区界面。新建仪表板后，该仪表板自动打开。仪表板界面与工作表界面类似，主要区别是左侧的“仪表板”窗口替代了工作表的“数据”窗格。“仪表板”窗口列出了当前在工作簿中的全部工作表，如图 5.150 所示。

添加工作表视图。单击工作表并按住鼠标左键将其从仪表板窗口拖至右侧的仪表板工作区中，拖动过程中出现的灰色阴影区域提示该工作表预计放置的位置，到达合适的位置后松开鼠标左键即可。注意，仪表板窗口中有勾选标记的工作表表明该工作表已经应用于仪表板，如图 5.150 所示显示

新建的仪表板包含 3 个工作表。

“布局”区域显示了仪表板包含的工作表和对象的布局方式，如图 5.151 所示。

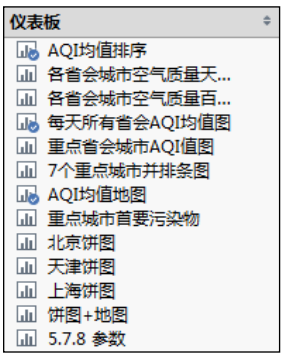


图 5.150　“仪表板”窗口

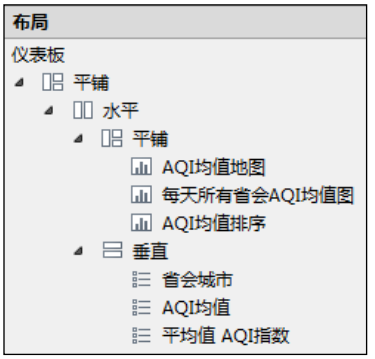


图 5.151　“布局”区域

布局包含“平铺”和“浮动”两种，如图 5.151 所示显示了 3 个工作表（包含 3 个图例）的布局方式。“平铺”是将仪表板中的工作表和对象放在一个布局容器中，工作表和对象自动调整大小以适应仪表板大小，如图 5.152 所示。“浮动”是将仪表板中的工作表和对象放在多层布局容器中，工作表和对象可以层叠在其他对象上面，如图 5.153 所示。

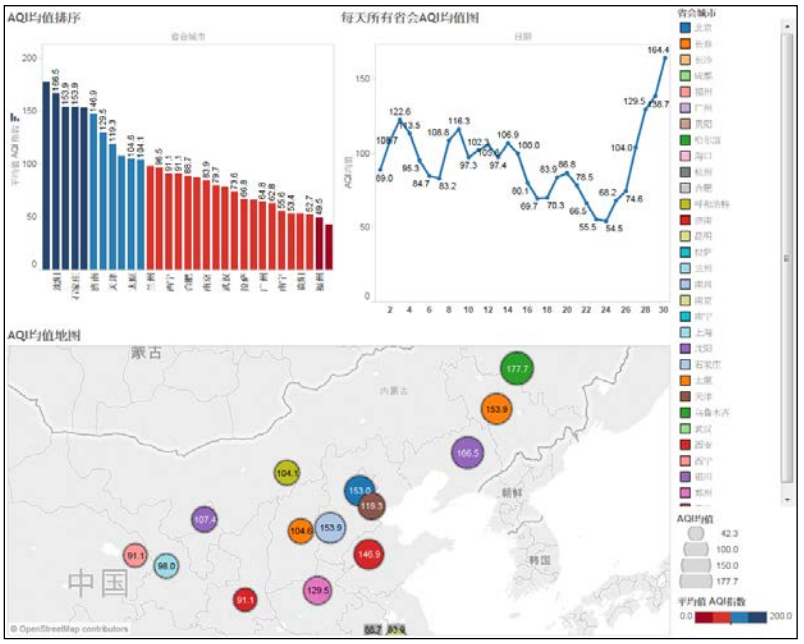


图 5.152　“平铺”布局

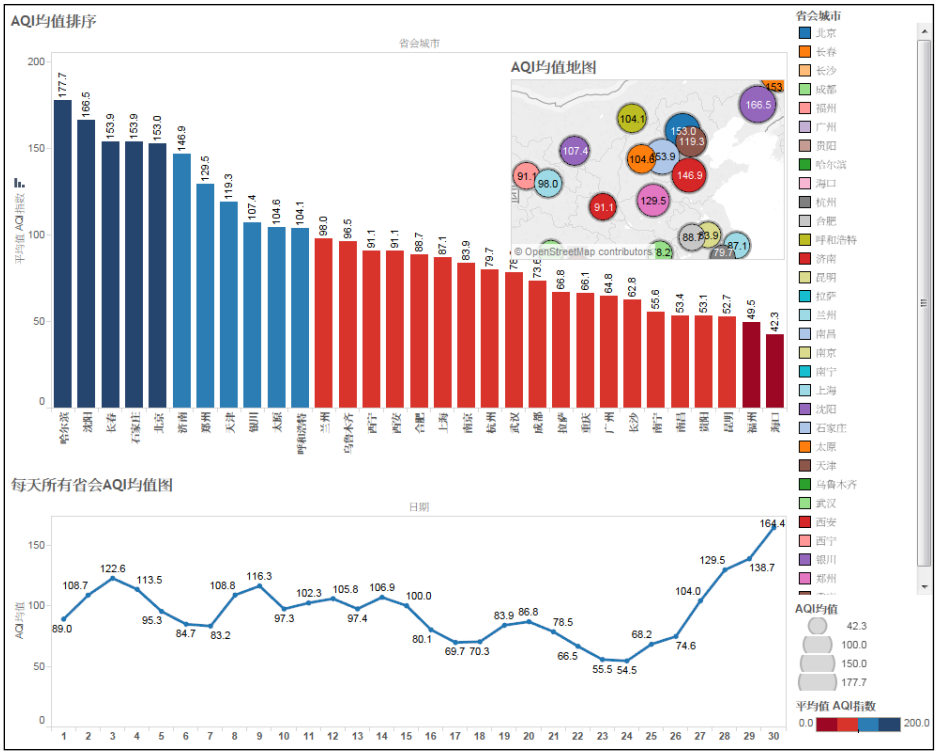


图 5.153 “浮动”布局

默认情况下，将仪表板设置为“平铺”布局。使用“浮动”布局往往是为合理利用仪表板中某个工作表或对象的空白部分。例如，图 5.153 中“AQI 均值地图”工作表浮动显示在工作表“AQI 均值排序”右上角的空白部分。工作表和对象的布局可以任意切换。如将图 5.152 中的 “AQI 均值地图”工作表切换为“浮动”布局，需要在仪表板中单击选中“AQI 均值地图”工作表，在仪表板左下角的“AQI 均值地图”窗格中勾选“浮动”复选项，如图 5.154 所示。拖动该浮动工作表到合适的位置，注意窗格中 *X* 值和 *Y* 值的变化，也可以手动调整该工作表的大小或在该窗格中设置宽和高。

图 5.154 “AQI 均值地图”窗格

添加其他仪表板对象。仪表板最主要的内容是工作表视图，也可以根据情况添加其他对象，如文本、图像、网页和空白区域，如图 5.155 所示。

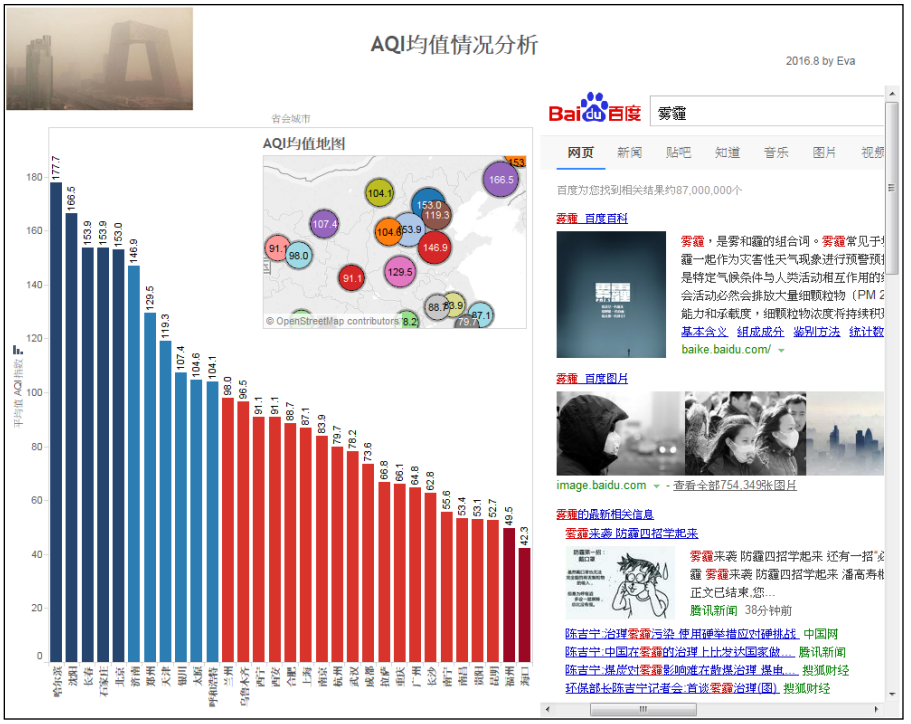


图 5.155　为仪表板添加其他对象效果

图 5.155 所示的仪表板顶部包含仪表板标题、图片和文本，左侧包含 2 个工作表视图，分别是“AQI 均值排序”（平铺）和“AQI 均值地图”（浮动），右侧包含一个网页。

单击“仪表板”窗口中的一个对象，设置对象的布局方式后即可为仪表板添加其他对象，如图 5.156 所示。

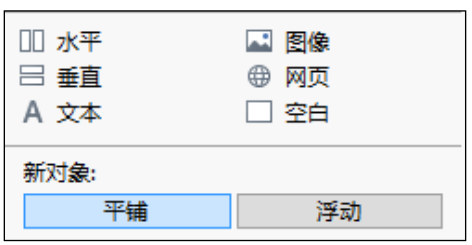


图 5.156　添加其他仪表板对象

注意，仅能为仪表板添加静态图像文件，如 JPG 或 PNG 等格式文件。如添加 GIF 格式的动态图片，仪表板仅显示该 GIF 图片的第一帧。为仪表板添加指定网页的 URL 时，必须在连接到 Internet 的前提下才能查看到嵌入到仪表板中的网页内容。注意，仪表板打印为 PDF 格式时，不显示网页的内容。

5.8.2 布局容器

布局容器方便用户在仪表板中组织工作表视图和对象。布局容器是仪表板中的一块区域，在此区域中，可以放置一个或多个工作表视图和对象，工作表视图根据容器中其他对象的大小和位置自动调整大小和位置。

添加布局容器。单击图 5.156 中左上方的“水平”或“垂直”布局容器按钮，然后将“水平”或“垂直”布局容器拖至仪表板工作区，可以添加一个“水平”或“垂直”布局容器，然后将工作表和对象添加到布局容器中。一个布局容器中既可以添加一个或多个工作表视图和对象，也可以添加一个或多个布局容器。

删除布局容器。打开选定对象右上角的下拉菜单，选择【选择布局容器】选项。例如，单击仪表板的网页对象，打开右上角的下拉菜单，选择【选择布局容器】选项，如图 5.157 所示。布局容器被选择后四周呈现蓝色框，单击右上角的三角形按钮，在打开的下拉菜单中选择【从仪表板移除】选项，如图 5.158 所示，该布局容器及其包含的对象将从仪表板移除。注意下拉菜单中的【移除容器】选项，表示该布局被移除，但容器中包含的对象依旧存在。

图 5.157 选择【选择布局容器】选项

图 5.158 选择【从仪表板移除】选项

设置布局容器的格式。可以设置布局容器的阴影和边框样式。默认情况下，布局容器透明且无边框。选择一个布局容器，用鼠标右键单击右上角的三角形按钮，打开下拉菜单，选择【设置容器格式】选项，如图 5.159 所示。在打开的“设置容器格式”窗口中设置“阴影”和“边界”，如图 5.160 所示。“阴影”用于设置颜色和不透明度，“边界”用于设置边框的线条样式、颜色和粗细。

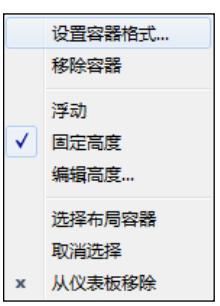


图 5.159 选择【设置容器格式】选项

图 5.160 设置容器格式

5.8.3　编辑仪表板

仪表板的命名、重命名、删除、复制和移动等操作与工作表的相关操作类似。

设置仪表板大小。根据展示仪表板的设备需求，用户可以设置合适的仪表板尺寸。使用“仪表板”窗口底部的区域可以设置仪表板的尺寸。默认情况下，仪表板尺寸是 1000×800 像素，单击“大小”文本框后面的下拉按钮，在打开的下拉列表中可以选择其他尺寸，如图 5.161 所示。

图 5.161　设置仪表板大小

仪表板大小的选项包含以下几种。

“自动”选项：表示仪表板自动调整大小，以填充应用程序窗口。

“精确”选项：表示仪表板一直保持固定大小。如果仪表板比窗口大，那么仪表板将使用滚动条滚动显示。

“范围”选项：表示仪表板在指定的最小和最大尺寸之间进行缩放，之后将显示滚动条或空白。

“预设”选项：表示从多种预先设定好的尺寸中选择，例如“便携式计算机”、“A2 横向”、“小型博客”或“iPad 纵向”等。若选择的预设尺寸比窗口大，则仪表板显示滚动条。

显示/隐藏工作表的组成部分。根据需要可以设置是否显示工作表的组成部分，如是否显示标题、说明、图例或快速筛选器。单击选定视图右上角的下拉按钮，或者在“仪表板”窗口中用鼠标右键单击“布局”区域中的一项，在弹出的快捷菜单中勾选要显示/隐藏的项即可，如图 5.162 所示。

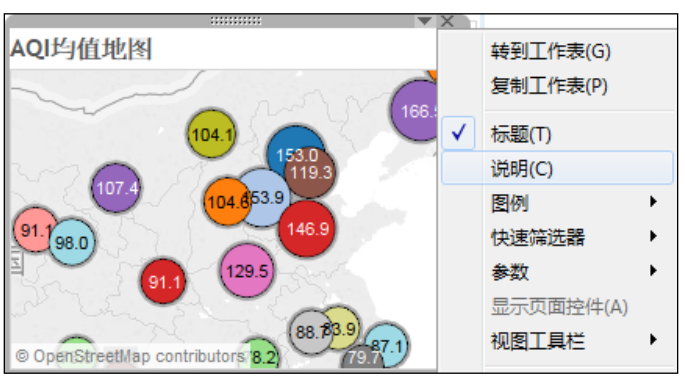


图 5.162　显示/隐藏工作表的组成部分

移动工作表视图/对象。根据需要可以重新排列仪表板中的工作表视图、图例、文本、网页或快速筛选器等。选择要移动的工作表视图或对象，单击视图或对象顶部的移动控制柄，将其拖到仪表

板的其他位置，可以放置该对象的位置显示为灰色阴影，如图 5.163 所示。

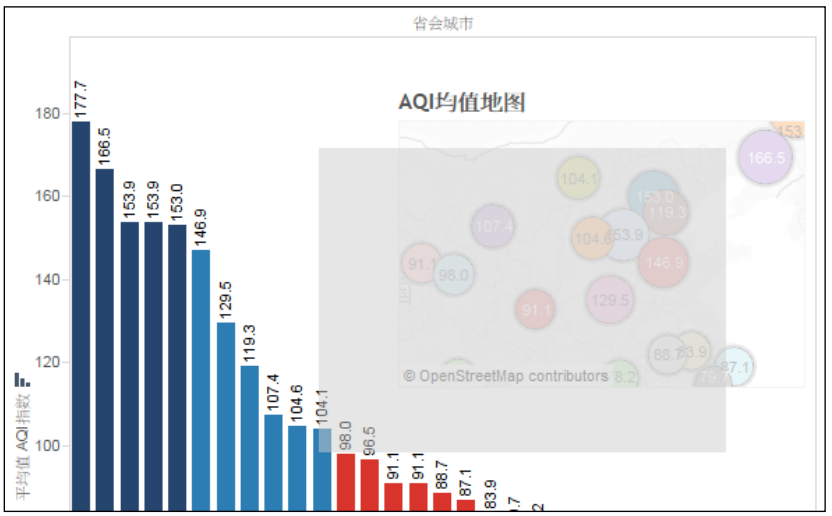


图 5.163　移动工作表视图/对象

移除工作表视图/对象。可以方便地从仪表板中移除工作表视图或对象。最常见且简单的方法是选择要移除的工作表视图或对象，单击右上角的“关闭”按钮。也可以选择要移除的工作表视图或对象，单击视图或对象顶部的移动控制柄，将其拖到仪表板工作区外。还可以在“仪表板”窗口中选择要移除的工作表，用鼠标右键单击该工作表，在弹出的快捷菜单中选择【从仪表板移除】选项，如图 5.164 所示。还可以在仪表板中选择要移除的视图或对象，在仪表板视图菜单中选择【从仪表板移除】选项，如图 5.165 所示。

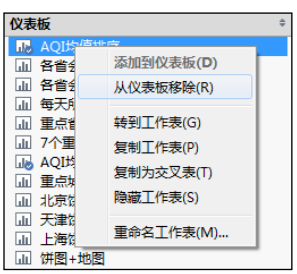


图 5.164　从“仪表板”窗口中移除工作表

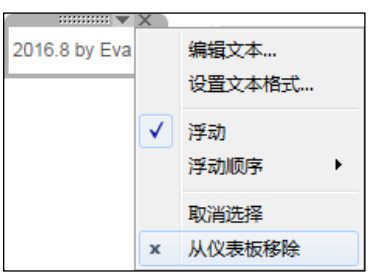


图 5.165　从仪表板视图菜单中移除对象

5.8.4　仪表板和工作表

仪表板是工作表的展示和汇总，二者互相更新，对工作表的修改会影响包含该工作表的仪表板，反之亦然。用户可以从仪表板转到工作表，方便编辑工作表。还可以从仪表板复制工作表、隐藏/显示仪表板中的工作表。

转到工作表。从仪表板可以快速地转到工作表视图，并对工作表进行编辑等操作。选择视图中的一个工作表，单击仪表板右上角的三角形按钮，在打开的下拉菜单中选择【转到工作表】选项，如图 5.166 所示。

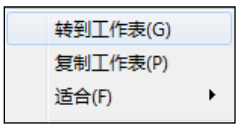


图 5.166　仪表板下拉菜单

复制工作表。可以在仪表板中复制一个与当前仪表板中的工作表完全相同的新工作表。选择视图中的一个工作表，单击仪表板右上角的三角形按钮，在打开的下拉菜单中选择【复制工作表】选项即可。

隐藏工作表。因为仪表板可以快速转到工作表，所以可以隐藏工作表，避免过多的工作表显示在标签栏上。用鼠标右键单击工作簿底部标签栏上的一个工作表，在弹出的快捷菜单中选择【隐藏工作表】选项，则该工作表被隐藏，工作表不会出现在标签栏上。

显示已经隐藏的工作表。虽然工作表被隐藏，但使用该工作表的仪表板并没有任何变化，而且可以使用【转到工作表】选项快速地从仪表板转到该工作表的编辑视图。

5.8.5　操作

Tableau 操作适合添加上下文和数据交互性，将分析结果直接链接到网页、文件或其他工作表。操作包含“筛选器”、“突出显示”和“URL”三种。如在显示利润的仪表板中，使用“筛选器”快速筛选符合条件的类别，并突出显示相关信息，同时使用“URL”打开产品网页。

1．筛选器操作

筛选器操作可以显示源工作表与其他一个或多个目标工作表之间的关联信息。源工作表和目标工作表通过一个共有字段相连，将这些工作表放入一个仪表板中展示时，源工作表通过共有字段筛选后，仪表板中其他目标工作表的显示内容将更新为仅显示相关信息。筛选操作特别适合数据量大时仅筛选用户感兴趣的内容。

编辑“库房图”工作表。打开 5.7.12 小节制作的工作表“库房图”，单击“标记”卡中的“颜色”，设置颜色为白色。单击“标记”卡中的“大小”，设置合适的大小。

创建“利润”工作表。新建工作表，将“订购日期”拖到“列”功能区，将“利润”拖到“行”功能区，展开“智能显示”，选择“面积图（连续）”。在“列”功能区上，单击“订购日期”下拉按钮，在下拉菜单中选择“月 2015 年 5 月”。将“邮寄方式”拖到“标记”卡中的“颜色”上，将“子类别”拖到“标记”卡中的“工具提示”上。重命名工作表为“利润”，最终效果如图 5.167 所示。

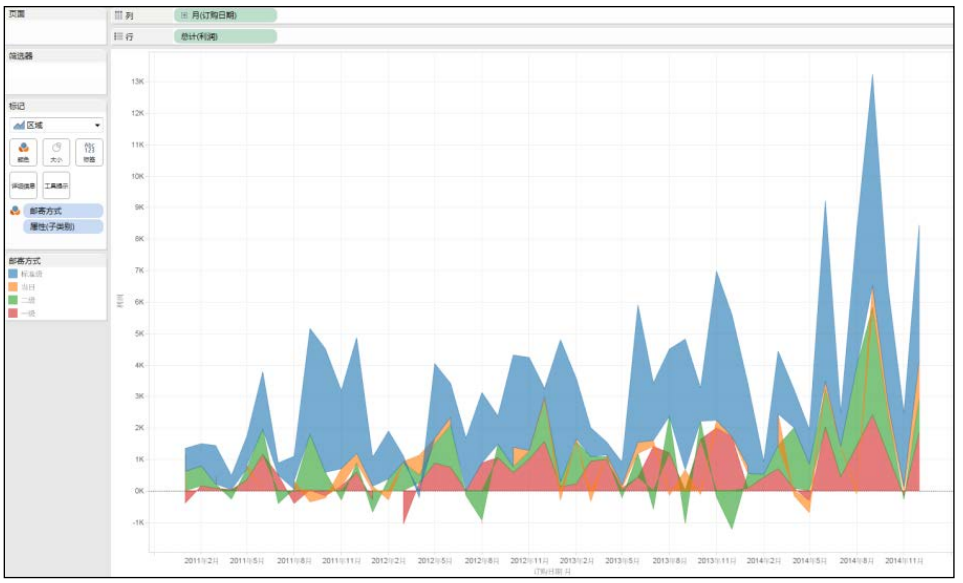


图 5.167 “利润”工作表

新建仪表板“筛选操作”。将“库房图”工作表和“利润”工作表以水平布局的方式拖入仪表板中，从仪表板中移除“邮寄方式”图例。重命名仪表板为“筛选操作”。

选择一个视图以用做筛选器。在仪表板中，单击“库房图”工作表右上角的下拉按钮，从下拉菜单中选择【用做筛选器】选项，如图 5.168 所示。

编辑操作。单击【仪表板】|【操作】选项，打开“操作”对话框。选择对话框中唯一的操作，如图 5.169 所示，然后单击“编辑”按钮。

图 5.168 选择【用做筛选器】选项

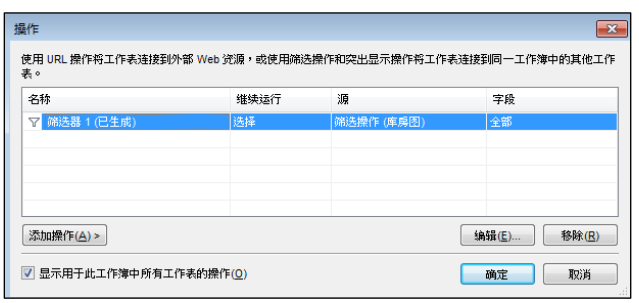


图 5.169 “操作”对话框

设置筛选器操作。在“编辑筛选器操作”对话框中，选择源工作表“库房图”，选择目标工作表“利润”，设置“运行操作方式”是“选择”，设置“清除选定内容将会”是“排除所有值”，如图 5.170

所示，然后单击“确定”按钮。。

“运行操作方式”选项决定如何启动操作，选项包含“悬停”、“选择”和“菜单”三种。

“悬停”选项表示通过将光标放置在视图的标记上运行操作，此选项适合在仪表板中突出显示和筛选器操作。“选择”选项表示通过单击视图中的标记运行操作，此选项适合所有类型的操作。“菜单”选项表示用鼠标右键单击视图中选定的标记，然后在快捷菜单中选择一个选项，此选项适合筛选器和 URL 操作。

“清除选定内容将会”选项决定在清除视图中的选项后执行的操作，包含“保留筛选器”、“显示所有值”和“排除所有值”三种。“保留筛选器”选项表示将筛选器保留在目标工作表上，仪表板上的目标视图将显示筛选结果。“显示所有值”选项表示将筛选器更改为包括所有值。“排除所有值”选项表示将筛选器更改为排除所有值。如果在另一个工作表中选择了值，可使用此选项生成只显示部分工作表的仪表板。

“目标筛选器”选项指定要在目标工作表中显示的数据。本例中选择的是筛选“所有字段”，也可以对“选定的字段”定义筛选器。若要定义特定字段的筛选器，可单击“添加筛选器”按钮进行添加。

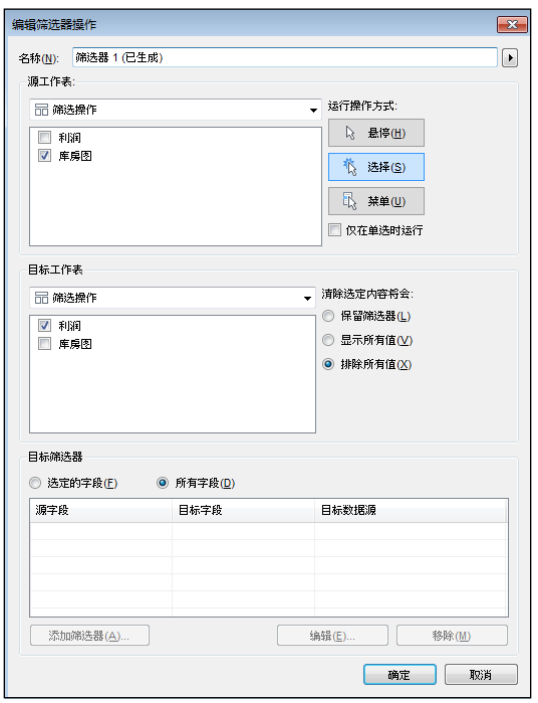


图 5.170　“编辑筛选器操作”对话框

查看筛选器操作效果。筛选操作设置完成后，“子类别”的选择将影响“利润”工作表，如单击仪表板中“库房图”中的“装订机”子类别，“利润”工作表将联动操作，如图 5.171 所示。

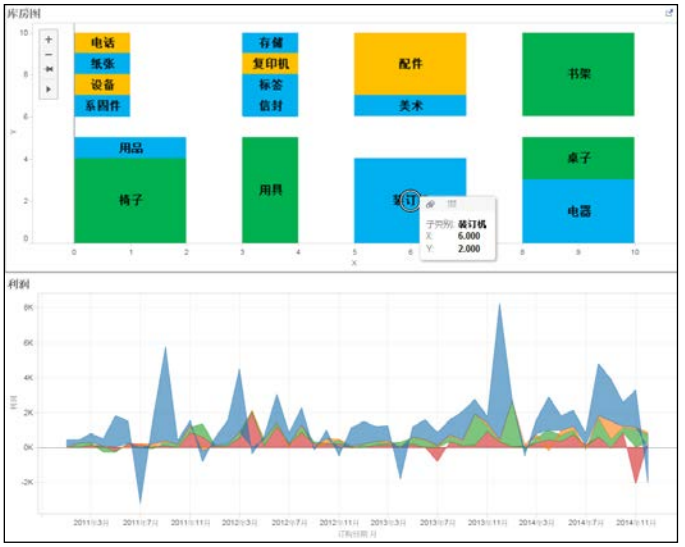


图 5.171 “筛选操作”仪表板

2．突出显示

使用突出显示操作可以为选择的标记添加颜色，其他所有标记显示为灰色，以突出对感兴趣的标记的关注。

选择突出显示的标记。在工作表中可以根据用户需要选择一个或多个标记进行突出显示，其他标记将自动显示为灰色。

打开工作表“利润”，复制该工作表为“利润气泡图”，然后使用“智能显示”将其修改为气泡图。按住键盘上的【Ctrl】键后用鼠标单击多个标记，如图 5.172 所示。也可以按住鼠标左键并移动到目标位置以框选多个标记，如图 5.173 所示。选择的多个标记呈现高亮显示，其他标记是灰色显示的。注意，工作簿保存后，下次打开该工作簿时选择的多个标记依旧呈现高亮显示。

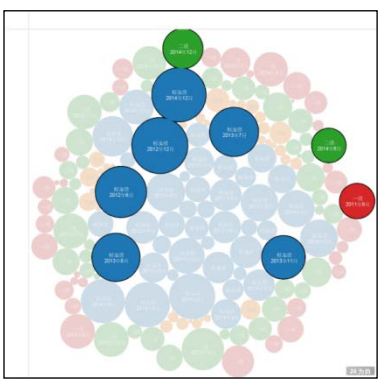


图 5.172 按住【Ctrl】键多选标记

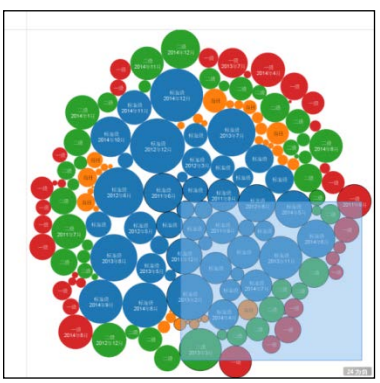


图 5.173 用鼠标框选多个标记

颜色图例突出显示。颜色图例突出显示启用后，与颜色图例中选定项关联的标记将着色，其他所有标记灰色显示。

打开“利润气泡图”工作表。单击“邮寄方式”图例右上角的三角形按钮，在打开的下拉菜单中勾选【突出显示选定项】，也可以单击下拉菜单顶部的突出显示按钮，如图 5.174 所示，效果如图 5.175 所示。注意启用颜色图例突出显示后，单击【工作表】|【操作】选项可以查看生成的突出显示操作。

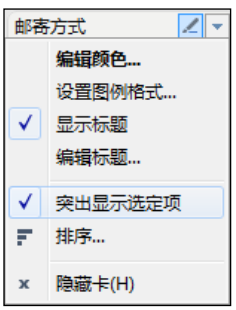


图 5.174　设置突出显示选定项

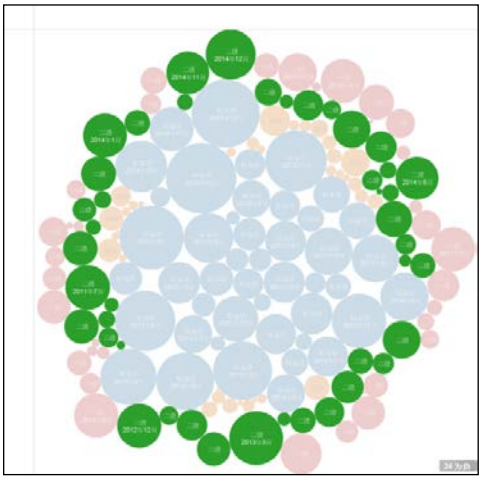


图 5.175　颜色图例突出显示

取消颜色图例突出显示。单击“邮寄方式”图例右上角的三角形按钮，在打开的下拉菜单中取消【突出显示选定项】的勾选。也可以再次单击下拉菜单顶部的突出显示按钮。注意该操作将从“动作”对话框中移除。

编辑突出显示操作。单击【工作表】|【操作】选项，选择生成的突出显示操作后单击“编辑”按钮，根据需要更改突出显示操作，如将“运行操作方式”修改为“悬停”。效果为将鼠标指针悬停时，颜色图例突出显示。

3．URL 操作

URL 操作是一种超链接，可以指向一个网页、文件或其他基于 Web 的资源。

添加 URL 操作。在工作表上单击【工作表】|【操作】选项，然后在仪表板上单击【仪表板】|【操作】选项，在打开的“操作”对话框中单击“添加操作”按钮，在弹出的下拉菜单中选择“URL”，如图 5.176 所示。

打开“添加 URL 操作”对话框，输入 URL 链接的名称，使用下拉列表选择源工作表或数据源，再设置运行操作方式，输入 URL 地址，如图 5.177 所示，单击两次“确定”按钮，关闭对话框并返回视图。

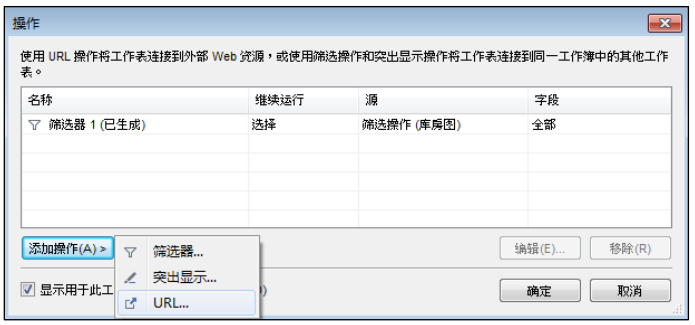


图 5.176　添加 URL 操作

图 5.177　“添加 URL 操作”对话框

在图 5.177 中，“URL 选项”选项区中有“URL 编码数据值”和“允许多个值”两个选择。勾选“URL 编码数据值”表示 URL 地址包含了 URL 中不允许使用的字符。勾选“允许多个值”表示 URL 地址中可能采用值列表作为参数。运行效果如图 5.178 所示，单击“超链接 1”可以进入百度首页。

在“URL”文本框中可以使用字段和筛选器值。例如，在百度首页搜索“test”的 URL 是“ https://www.baidu.com/s?wd=test&rsv_spt=1&rsv_iqid=0xd7731d1d0013dd9c&issp=1&f=8&rsv_bp=1&rsv_idx=2&ie=utf-8&tn=baiduhome_pg&rsv_enter=1&oq=%E6%89%93%E5%8D%B0%E6%9C%BA&inputT=1232&rsv_t=ec78edcogt4bUIySx4Q3VyO5nqFefGtLwyXb%2Btg5agmgZmaHVNDnh88riPbXuzt%2FNLtv&rsv_pq=f62bc478001cfeb2&bs=%E6%89%93%E5%8D%B0%E6%9C%BA”，可以使用字段和筛选器值替代要搜索的内容，如图 5.179 所示。

图 5.178　运行效果

图 5.179　URL 中使用的字段

5.9 故事

故事包含多个呈现相关信息的工作表和仪表板。使用故事可以完整地展示工作表和仪表板之间的关系，帮助用户演示决策与结果的关系。故事可以保存在本地计算机，也可以发布到网络共享给其他用户。

故事中包含的多个工作表和仪表板按顺序排列，每个工作表或仪表板包含一个说明，也称“故事点”。故事中的每个“故事点”连接至工作表和仪表板，所以故事呈现的内容会在更改工作表或仪表板后实时更新，对仪表板进行的更改也会影响其包含的工作表和仪表板。因此，故事是动态的。故事发布后，用户可以使用筛选器与故事互动。

新闻记者或编辑可以使用故事分析数据背后的新闻真相，显示数据随时间变化的效果，也可以实现假设分析，分析新闻可能带来的其他问题。也可以用故事讲述一个新闻事件，方便受众的理解。

创建故事。单击【故事】|【新建故事】选项，也可以单击标签栏中的“新建故事”图标，工作表底部的标签栏处会显示新建的故事标签，默认名字“故事 1”。为第一个故事点添加说明，将工作表或仪表板拖到故事中并放在视图的中央，然后单击“新空白点”按钮添加新故事点。

故事工作区界面。新建故事后该故事自动打开。故事界面左侧是“仪表板和工作表”窗格，右侧是故事标题和故事导航器。

“仪表板和工作表”窗格上方显示可以使用的工作表和仪表板，下方是“说明”、导航器是否显示设置和故事大小设置，如图 5.180 所示。

“说明”选项用于向故事点添加说明。在“仪表板和工作表”窗格中双击或拖动“说明”到故事工作区可以添加说明，一个故事点可以添加多个说明。“说明”可放置到故事工作区的任意位置，“说明”仅存在于故事点上，但不会影响原始工作表或仪表板。“说明”添加后出现一个描述框，可以选

择并将其移动到合适的位置。选择“说明”时，单击下拉按钮后打开下拉菜单，可以编辑说明、设置说明格式、设置其相对于它可能覆盖的其他任何说明框的浮动顺序或取消选择说明，或从故事点将其移除，如图 5.181 所示。

是否勾选“显示前进/后退按钮”复选框决定故事导航器的显示效果，效果如图 5.182 所示。

设置故事大小。根据展示故事的设备需求，可以设置合适的故事尺寸。默认情况下，故事尺寸是 1016×964 像素，单击“大小”下拉列表框中的下拉按钮，在下拉列表中可以选择其他尺寸，以适合故事里的工作表和仪表板。

图 5.180 “仪表板和工作表”窗格下方

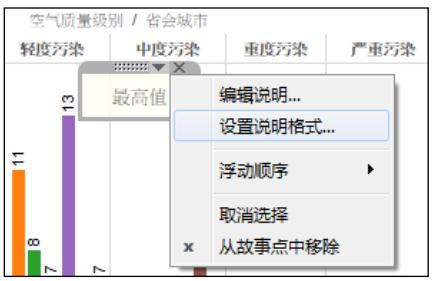


图 5.181 故事“说明”下拉菜单

故事标题和故事导航器如图 5.182 所示。导航器用来显示、编辑和组织所有故事点。故事导航器包含“新空白点”和“复制”两个按钮。“新空白点”按钮用于添加一个新的空白故事点，“复制”按钮用于复制当前故事点。

图 5.182 故事标题和故事导航器

编辑故事格式。单击【故事】|【设置格式】选项，打开“设置故事格式”窗格，可以编辑故事阴影、标题、导航器和说明，如图 5.183 和图 5.184 所示。

其中，“故事阴影”选项区用于设置故事是否有阴影、阴影的颜色和透明度。“故事标题”选项区用于设置标题的字体、对齐方式、阴影和边界效果。“导航器”选项区用于设置导航器的字体和阴影。“说明”选项区用于设置说明的字体、对齐方式、阴影和边界效果。单击“设置故事格式”窗格底部的“清除”按钮，可以清除所有格式设置。

演示故事。故事制作完成后，可以进入演示模式演示故事。单击工具栏中的“演示模式”按钮或按【F7】键进入演示模式，再次单击“演示模式”按钮或按【Esc】键或【F7】键可退出演示模式。

图 5.183　设置故事阴影和标题

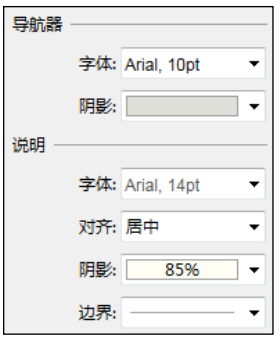


图 5.184　设置故事导航器和说明

编辑故事点。故事点可以被删除、更新或重新排序。在导航器中单击要编辑的故事点，单击右上方的删除图标可以删除该故事点。选中故事点后按住鼠标左键拖动该故事点到合适的位置，也可以双击故事点以更新故事点的内容。

5.10 作品发布

Tableau Public[1]是Tableau公司提供的一个将交互数据发布到Web的免费服务。任何人均可使用Tableau Public与数据交互、下载数据或创建自己的数据可视化项。保存到Tableau Public的工作簿的数据不得超过 100 万行。

5.10.1　工作簿和工作表

Tableau 使用了工作簿和工作表文件结构，这与 Microsoft Excel 十分类似。Tableau 包含多种文件类型，如工作簿文件、书签文件、打包数据源文件、数据提取文件和数据连接文件。

工作簿文件的扩展名为.twb。工作簿中含有一个或多个工作表、仪表板和故事，其中仪表板和故事不是必须包含的内容。工作表包含单个视图、功能区、图例和“数据”窗格。仪表板是多个工作表中的视图的集合，故事是多个工作表和仪表板的集合。

书签文件的扩展名为.tbm。书签包含单个工作表，是快速分享所做工作的简便方式。单击【窗口】|【书签】|【创建书签】选项，在打开的“创建书签”对话框中，指定书签文件名和位置后可以创建书签。书签文件默认的保存位置是 Tableau 存储库的“Bookmarks”文件夹。

打包工作簿文件的扩展名为.twbx。打包工作簿是一个压缩文件，包含一个工作簿及任何支持本地文件的数据源和背景图像。这种文件适合于不能访问该数据的他人查看。

1　https://public.tableau.com/。

数据提取文件的扩展名为.tde。数据提取文件是部分或整个数据源的一个本地副本，可用于共享数据、脱机工作并提高数据库性能。

数据源文件的扩展名为.tds。数据源文件是快速连接到经常使用的数据源的快捷方式。数据源文件不包含实际数据，而只包含连接到数据源所必需的信息和在“数据”窗格中所做的修改，如默认属性或计算字段等。

打包数据源文件的扩展名为.tdsx。打包数据源是一个压缩文件，包含数据源文件（.tds）及任何本地文件数据源，例如数据提取文件（.tde）、文本文件、Excel 文件、Access 文件和本地多维数据集文件。此文件可以与无法访问计算机本地存储的原始数据的他人分享。

以上文件保存在“我的 Tableau 存储库”目录中的关联文件夹中。可以更改存储库位置为任意一个文件夹。单击【文件】|【存储库位置】选项，在打开的“选择存储库文件夹”对话框中选择一个新文件夹作为新的存储库位置，如图 5.185 所示，然后单击“选择文件夹”按钮，重新启动 Tableau 则使用新存储库。

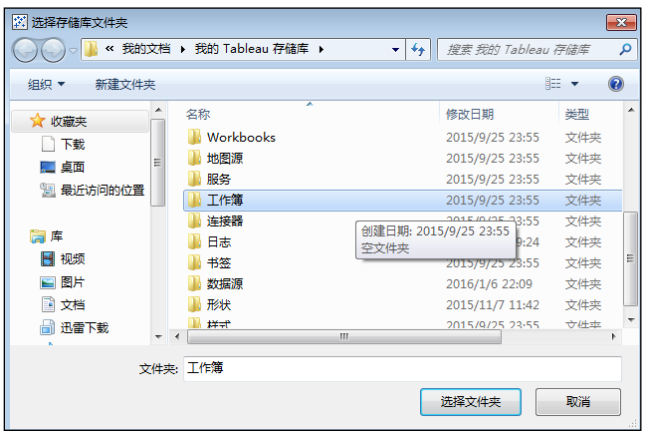


图 5.185 更改存储库位置

5.10.2 发布

个人用户使用的 Tableau Professional 版本必须作为用户添加到 Tableau Server 并被授予发布权限后，才能使用 Tableau Professional 发布视图和数据源。而 Tableau Personal 版本用户只能发布到 Tableau Public。将工作簿保存到 Tableau Public 的操作步骤如下。

（1）单击【服务器】|【Tableau Public】|【保存到 Tableau Public】选项。

（2）使用个人账号信息登录 Tableau Public，如图 5.186 和图 5.187 所示。如果没有账号，可以免费创建。注意，只有登录账号后才能共享、下载、删除和上传 Tableau 作品。

图 5.186　个人账号登录

图 5.187　登录后显示个人信息

（3）显示已发布的工作簿，用户可以预览所有保存的工作表。选择一个工作表并单击视图左下角的“共享”按钮以获得一个链接，如图 5.188 所示。用户可将此链接通过电子邮件发送到网页或嵌入到网页中方便他人使用。

图 5.188　链接信息

5.10.3　打印

Tableau 作品制作完成后，可将这些视图打印出来。首先使用“页面设置”对话框指定打印页面的外观，然后就可以打印者或发布为 PDF 文件。

“页面设置”对话框中的“常规”选项卡可以设置打印时显示的元素，如显示或隐藏标题、查看、说明、颜色图例、形状图例、大小图例和地图图例，如图 5.189 所示。

“页面设置”对话框中的“布局”选项卡可以设布局图例、页边距和居中选项。如果包括图例，可以为打印页面上的图例显示方式选择一个选项。“边距”选项区用于设置上、下、左、右的边距。还可以在页面上将视图设置为水平或垂直居中，如图 5.190 所示。

使用“打印缩放”选项卡，可将视图调整到特定尺寸或更改页面方向，如图 5.191 所示。“打印缩放”选项区用于设置缩放以使其适合一个页面或跨多个页面缩放。“页面方向”选项区用于设置视图在打印页面上的方向。其中，“使用打印机设置”表示使用打印机指定的页面方向；“纵向”表示垂直显示；“横向”表示水平显示。

图 5.189 “常规”选项卡

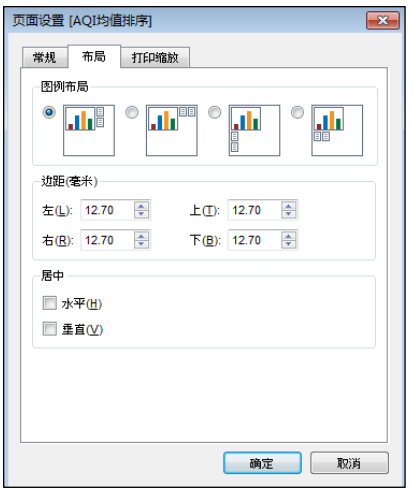


图 5.190 “布局”选项卡

“页面设置”对话框设置完成后，单击【文件】|【打印】选项进行打印，也可以打印为 PDF 文件。

首先单击【文件】|【打印为 PDF】选项，然后在打开的“打印为 PDF”对话框中选择打印范围，如图 5.192 所示。

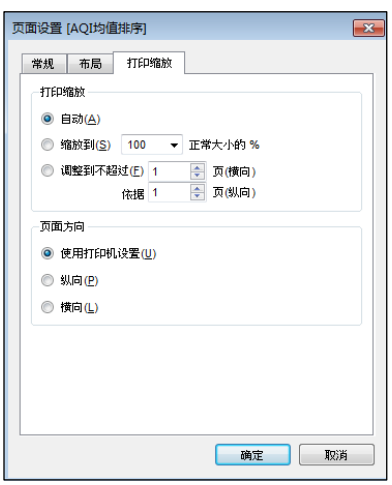


图 5.191 “打印缩放”选项卡

图 5.192 “打印为 PDF”对话框

然后选择“纸张尺寸”。如果希望在创建后自动打开 PDF 文件，则勾选“打印后查看 PDF 文件”复选框。勾选“显示选定内容”复选框，则视图中的选定内容将保留在 PDF 中。

单击“确定”按钮并指定保存位置，最后单击“保存”按钮完成保存操作。

5.11 Tableau 作品

Tableau Public 为用户提供了一个创建交互式的图表、地图、仪表板和应用的公共区域。用户的作品可以发布到 Tableau Public，方便用户互相学习。Tableau Public 上的优秀作品适合初学者学习，也值得专业制作人互相鉴赏。本节选取三个作品，展示 Tableau 的呈现效果。

5.11.1 Is Your Country Good at Reducing CO2 Emissions

该作品的中文名称是“你们国家的二氧化碳排放量是多少”，可以在“https://public.tableau.com/s/gallery/your-country-good-reducing-co2-emissions”查看，作者是 Yvan Fornes，发布于在巴黎举办的第 21 届联合国气候变化大会（COP21）召开前夕，作品呈现了各个国家或地区的二氧化碳总排放量与人口和 GDP 的关系，并使用工具计算每个人必须减少的二氧化碳排放量，以实现全球 2050 年的目标。

该作品的数据来源于世界银行。作品是动态的，包含四个部分。第一部分展示了全球的二氧化碳来自哪个国家和地区。使用矩形框展现 2011 年二氧化碳总排放量，二氧化碳总排放量前 5 名的国家的矩形框中间显示该国家的国旗，其面积的大小与二氧化碳排放量成正比。将鼠标指针移动到该矩形框时，以千万吨为计量单位显示该国家或地区的二氧化碳总排放量和排名，以吨为计量单位显示该国家或地区的人均二氧化碳排放量和排名，并用颜色区分人均二氧化碳排放量，如图 5.193 所示。以中国为例，二氧化碳总排放量是世界第 1 名，但人均二氧化碳排放量列世界第 49 名。

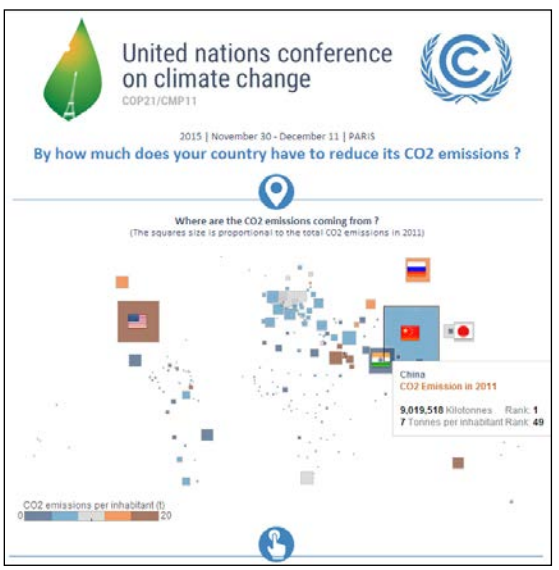


图 5.193　第一部分

向下滚动鼠标到第二部分，该部分展示“我们怎样才能改善这种情况”，选择国家并设置该国家二氧化碳排放目标，可以看到，为配合全球 2050 年目标，我们每个人必须减少的二氧化碳排放量。如在左侧蓝色文本框中输入“China”，在右侧蓝色方框中左右拖动指针可以设置减排目标为“–57%”，下方显示 2011 年和 2050 年人均二氧化碳排放量，即中国二氧化碳排放量下降比例，如图 5.194 所示。第二部分使用了筛选器，方便用户选择感兴趣的国家，而且 2050 年目标的设置简单易懂。

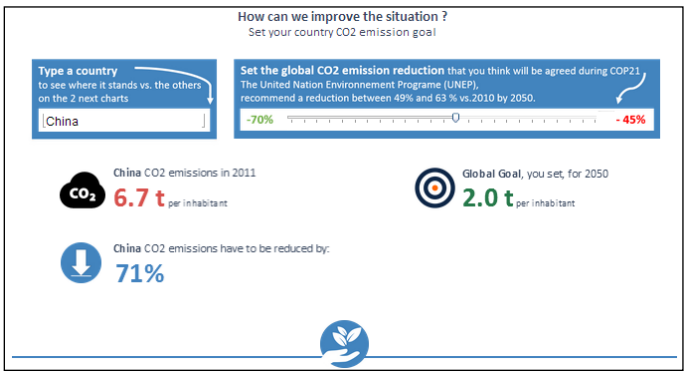


图 5.194　第二部分

向下滚动鼠标到第三部分，该部分展示了二氧化碳总排放量前五名的国家与其他国家或地区的对比。排在前五名的国家分别是中国、美国、印度、俄罗斯和日本。气泡大小正比于 2011 年二氧化碳排放总量。该部分还展示了人口与二氧化碳排放总量的关系，如图 5.195 所示，使用散点图并使用人均二氧化碳排放量平均线（2011 年全球人均二氧化碳排放量是 4.6）划分国家和地区。

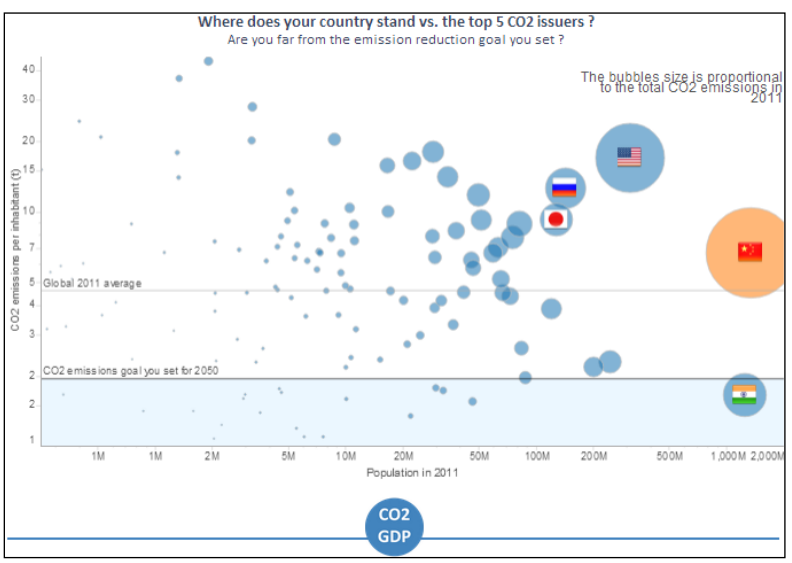


图 5.195　第三部分

继续向下滚动鼠标到第四部分，该部分展示了二氧化碳排放量与 GDP 的关系，该部分选取了十个人口最多的国家，呈现 1960 年至今二氧化碳排放量与 GDP 的关系图，如图 5.196 所示。

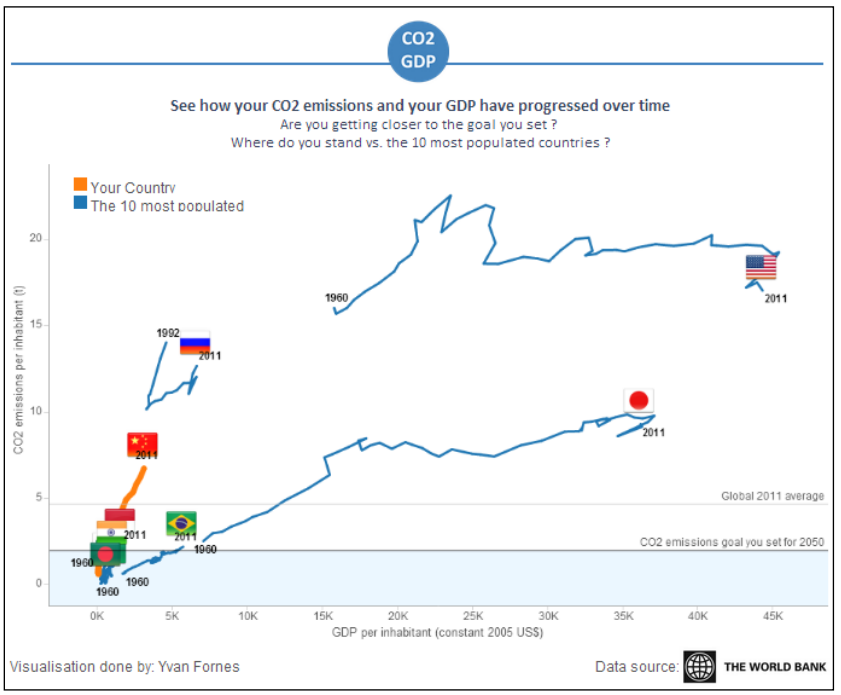


图 5.196　第四部分

在作品的右下角单击“下载”按钮可以下载为 TWB 格式、PDF 格式或图片格式的文件。也可以在“https://public.tableau.com/workbooks/2015UnitedNationsClimateChangeConferenceCOP21.twb”直接下载作品的 TWB 格式。

5.11.2　Cabs in NYC

该作品的中文名称是“纽约市出租车”，可以在“https://public.tableau.com/profile/adrien.charles#!/vizhome/Taxi-F1/Tableaudebord2”查看，作者是 Adrien Charles。该作品最初发布于其个人网站“http://www.adriencharles.com/”，呈现了纽约市 3 天内出租车路线图，使用了超过 5 百万个点用于用户导航，显示了白天特定时间出租车的活动情况，在底部单击“小时”即可查看出租车在这个特定时间的活动情况，如图 5.197 所示。

该作品的数据来源于 NYC Taxi and Limousine Commission（纽约出租车和轿车委员会），可以在“https://archive.org/details/nycTaxiTripData2013”下载。

该作品是动态的，包含 11 个工作表，由 4 部分组成。左上角展示了人们上下出租车的地点，蓝色表示上车地点，黄色表示下车地点。左侧是统计数据，右侧地图可以用鼠标滚轮放大，也可以单

击自己感兴趣的地点使其高亮显示。

左下角展示了人们乘坐出租车出行的时间，能够使用户方便地查看高峰出行时段并了解该时段出租车的平均车速（单位是英里/小时）及每次乘坐的时间，蓝色表示乘坐时间段，橘色表示乘坐时长（本书黑白印刷，建议上网查看本作品）。可以单击感兴趣的时间点，高亮显示相应信息。

右上角展示了人们从帝国大厦出发乘坐出租车的情况，一半左右的乘客的路程在 12 英里之内。右下角展示了人们从肯尼迪机场出发乘坐出租车的情况。

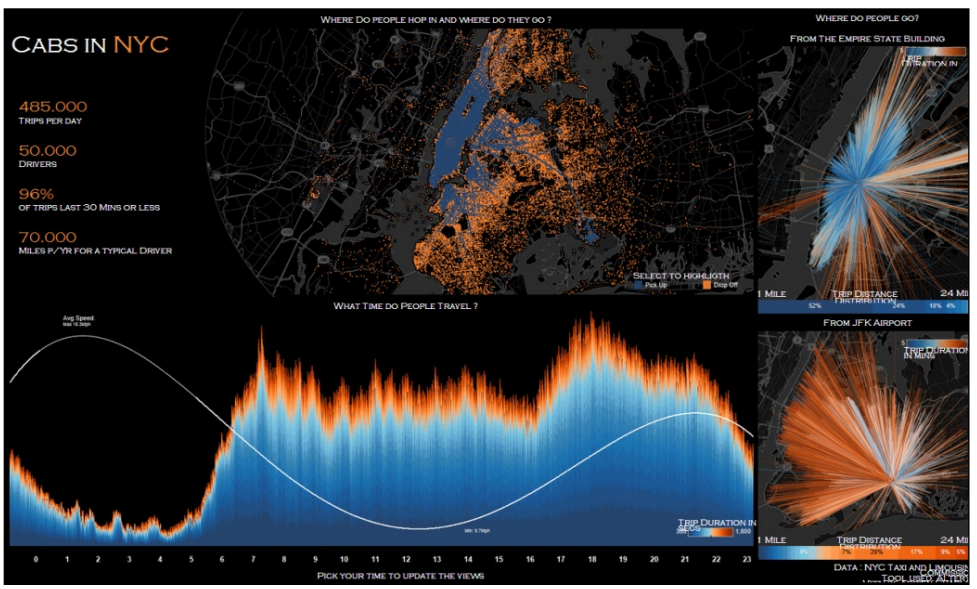


图 5.197　Cabs in NYC

5.11.3　Analysis of Twitter Hashtags Following the Paris Attacks

该作品的中文名称是“巴黎恐怖袭击后的 Twitter 主题标签分析”，可以在“https://public.tableau.com/s/gallery/analysis-twitter-hashtags-following-paris-attacks”查看，作者是 Jonathan Trajkovic，最初发布于“http://tipsandviz.blogspot.fr/”。这个可视化作品使用聚合 Twitter（推特）数据来描绘巴黎恐怖袭击后的反应，尝试告诉用户 Twitter 数据背后的故事，使用左右导航箭头展开仪表板可以查看连续的数据故事。

该作品的数据来自网站“http://www.talkwalker.com”。

该作品是动态的，包含 11 个故事点。故事点“6 hashtags”呈现了 6 个主题标签，即“PrayForParis”（为巴黎祈祷）、“ParisAttacks”（巴黎攻击）、“Fusillade”（扫射）、“RechercheParis”（巴黎研究）、“PorteOuverte”（巴黎大门开了）和“PrayForSyria”（为叙利亚祈祷）推特数量及时间的关系，如图 5.198 所示。使用面积图呈现推特数量比例，详细信息中包含时间点和比例，简单明了。

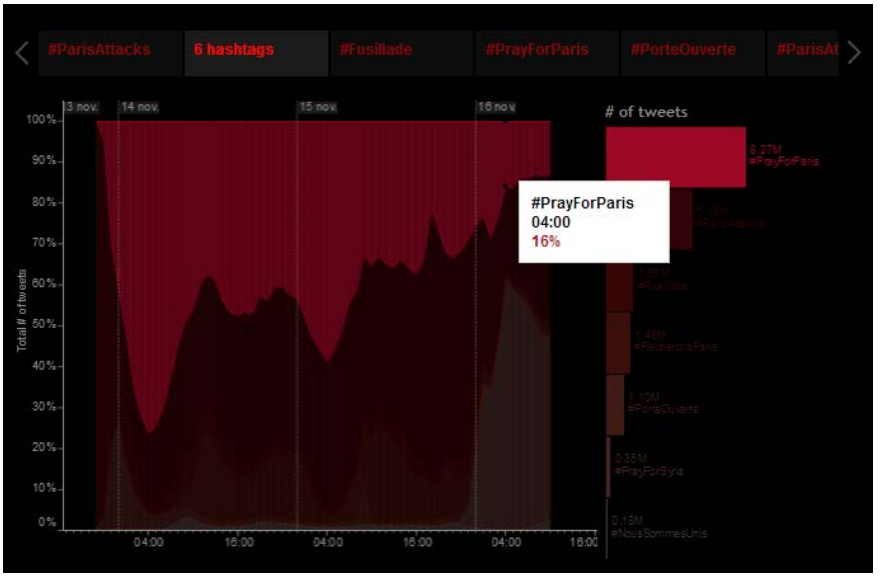


图 5.198　故事点“6 hashtags”

故事点“PrayForParis”（为巴黎祈祷）呈现了主题标签“PrayForParis”（为巴黎祈祷）在不同时间点的推特数量，并标明最高点峰值，如图 5.199 所示。

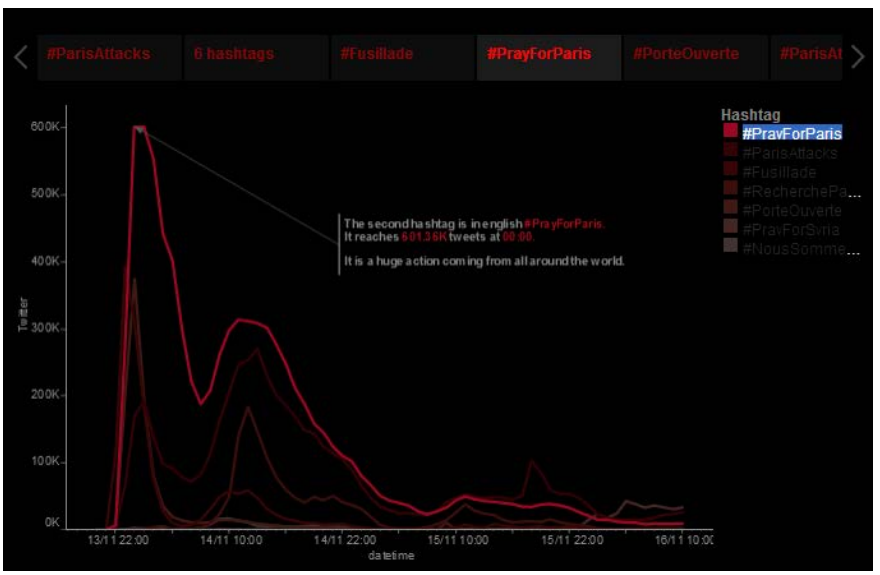


图 5.199　故事点“PrayForParis”

故事点“Millons of tweets”（无数的推特）呈现了在不同时间点的推特总数量，如图 5.200 所示。

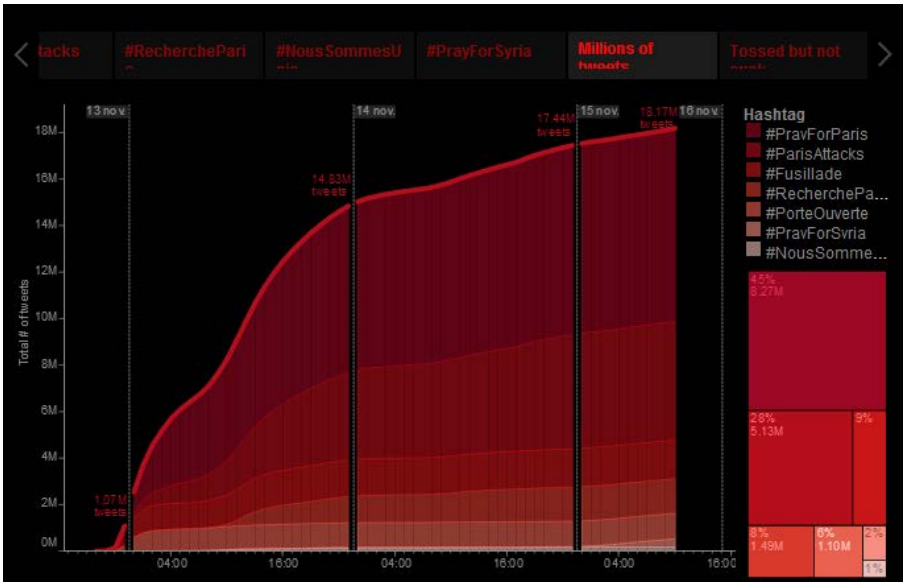


图 5.200　故事点“Millons of tweets”

在作品的右下角单击“下载”按钮可以下载该作品，也可以在“https://public.tableau.com/workbooks/ParisAttacks-HowTwittertellsthestory.twb”直接下载作品的 TWB 格式。

第 6 章

其他数据新闻制作工具

- 图表绘制工具库 ECharts
- 标签云
- 关系图制作工具 PeoplePlotr
- 语义万维网服务 Open Calais
- HTML5 网站制作模板

数据可视化工具 Tableau 并不是制作数据新闻的唯一工具，随着数据新闻包含的媒体种类越来越多，客户对数据新闻的要求和期待也在提升，为达到最佳的视觉效果，制作数据新闻时会根据需要使用很多其他类型的工具。

6.1 图表绘制工具库 ECharts

财新传媒 CTO 黄志敏说：“ECharts 是我接触过的最优秀的可视化工具，也是进步最快的软件，希望它早日成为世界级的开源项目。”

ECharts[1]是Enterprise Charts的缩写，是商业级数据图表，设计的最初目的是满足百度公司商业体系里各种业务系统的报表需求。

ECharts 本身是一个 JavaScript 图表库。2013 年 6 月 30 日，ECharts 发布了 1.0 版本，目前的最高版本是 3.2（截止到 2016 年 7 月）。ECharts 可以运行在 PC 平台上，也可以运行在移动设备上，它设计快速，能提供直观、生动、可交互且个性化的数据可视化图表。

ECharts 与 Tableau 均可实现数据的可视化，但二者的区别很大。其中最重要的区别就是 Tableau 不需要编写任何代码，最复杂的数据可视化就是添加或设置一些函数，而 ECharts 需要修改代码，甚至是编写代码。在数据的可视化呈现上两个工具各具特色。

6.1.1 获取 ECharts

用户可以通过多种方式获取ECharts。最常见的方法是在ECharts官网下载界面 [2]中选择版本进行下载。以 3.2.0 版本为例，下载界面中共有四个选项，其中“常用”选项包含常用的图表组件，“精简”选项只包含最基本的折线图、柱形图和饼图等，“完整”选项包含所有图表组件，“源代码”选项不仅包含所有图表组件，还包含源代码。

用户可根据需求个性化下载，建议初学者下载“完整”版本或“源代码”版本。“完整”版本下载的文件名是“echarts.min.js”，“源代码”版本下载的文件名是“echarts.js”。

文件无需解压，也无需安装，使用 ECharts 只需要像普通的 JavaScript 库一样用“<script>”标签引入即可。如引用“完整”版本 ECharts 的代码是：

```
<script src="echarts.min.js"></script>
```

然后就可以像其他的 JavaScript 库一样使用了。

6.1.2 绘制一个简单的图表

DOM 是 Document Object Model 的缩写，中文翻译为文档对象模型，DOM 以一种独立于平台和

1 官方网站 http://echarts.baidu.com。

2 下载网址 http://echarts.baidu.com/download.html。

语言的方式访问、修改一个文档的内容和结构。DOM 设计是以对象管理组织（OMG）的规约为基础的，因此可以用于任何编程语言。例如，通过 JavaScript 代码对 HTML 和 XML 数据进行 DOM 方式的操作，从而做到页面的动态修改、更新和数据的提取处理等。

首先，引入 ECharts 后，需要为 ECharts 准备一个具备高和宽的 DOM 容器。代码是：

```
<div id="main" style="width: 600px; height:400px;"></div>
```

标签用于把文档分割为独立的、不同的部分。一般用做严格的组织工具，并且不使用任何格式与其关联。

然后，通过 echarts.init( )方法初始化一个 ECharts 对象实例。

最后，使用 setOption( )方法生成一个简单的柱状图。

完整代码[1]如图 6.1 所示，代码来源于ECharts官网，笔者添加了注释，方便读者学习和理解。

```
<!-- <!DOCTYPE>声明用于指示Web浏览器关于页面使用哪个HTML版本进行编写的指令。 -->
<!DOCTYPE html>
<!-- 告知浏览器其自身是一个HTML文档。 -->
<html>
<!-- 此标签用于定义文档的头部，它是所有头部元素的容器。 -->
<head>
    <!-- charset属性规定此HTML文档的字符编码。 -->
    <meta charset="utf-8">
    <!-- 此标签定义文档的标题。 -->
    <title>ECharts</title>
    <!-- 引入 echarts.min.js -->
    <script src="echarts.min.js"></script>
</head>
<body>
    <!-- 为ECharts准备一个具备大小（宽高）的Dom -->
    <div id="main" style="width: 600px;height:400px;"></div>
    <!-- type 属性规定脚本的MIME类型。JavaScript的MIME类型是"text/javascript"。 -->
    <script type="text/javascript">
        // 基于准备好的DOM，初始化ECharts实例
        var myChart = echarts.init(document.getElementById('main'));
        // 指定图表的配置项和数据
        var option = {
            title: {
                text: 'ECharts 入门示例'    //设置图表的主标题文本
            },
            tooltip: {},                    //设置图表的提示框组件,默认显示提示框
            legend: {
                data:['销量']               //设置图表的图例组件的数据
            },
            xAxis: {                        //设置图表X轴的类目数据
                data: ["衬衫","羊毛衫","雪纺衫","裤子","高跟鞋","袜子"]
            },
            yAxis: {},                      //可在此处设置图表的Y轴
            series: [{                      //设置系列列表
                name: '销量',               //设置图表系列的名称
                type: 'bar',                //设置图表类型是柱形图
                data: [5, 20, 36, 10, 10, 20]//设置系列数据
            }]
        };
        // 使用刚指定的配置项和数据显示图表。
        myChart.setOption(option);
    </script>
</body>
</html>
```

图 6.1　绘制一个简单图表的源代码

1　源代码来源于 http://echarts.baidu.com/tutorial.html#5%20%E5%88%86%E9%92%9F%E4%B8%8A%E6%89%8B%20ECharts。

EditPlus 是一款功能强大的文本编辑器，编辑代码时它支持颜色标记、HTML 标记等功能，还能内建完整的 HTML 和 CSS 标签功能。初学者也可以使用“记事本”工具编写简单的代码。

将图 6.1 所示的源代码保存为 HTML 文档 “test.html”，并确保该文件与“echarts.min.js”在同一个文件夹下，使用浏览器打开“test.html”，查看图表效果，如图 6.2 所示。

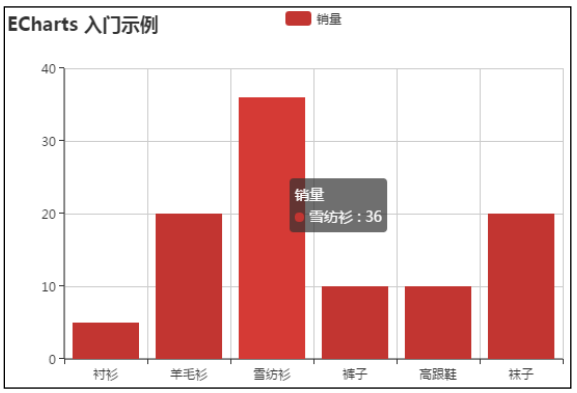


图 6.2　绘制的一个简单图表效果

为保证浏览效果，建议将浏览器升级到最高版本以支持 HTML5。也可以通过 http://www.w3school.com.cn/html5/html_5_video.asp 页面测试用户浏览器是否支持 HTML5。

6.1.3　编辑图表

编辑图表需要了解ECharts的功能，具体内容可参考ECharts的API[1]和配置项手册 [2]文档。本小节仅对图表进行基本的类型修改、增加图表显示内容及使用主题等操作。

- **修改图表类型**

ECharts 可以实现多种图表的制作，如常规的折线图、柱状图、散点图、饼图、地图、热力图、关系图、漏斗图和仪表盘等，并且支持图与图之间的混搭。

修改 series 中的 type 属性可以编辑图表类型，如“line”表示折线图，折线图是用折线将各个数据点标志连接起来的图表，用于展现数据的变化趋势；如“pie” 表示饼图，适合表现不同类目的数据在总和中的占比关系。下面的代码修改了图表类型，效果如图 6.3 所示。代码如下：

```
series: [{                          //设置系列列表
    name: '销量',                   //设置图表系列的名称
    type: 'line',                   //设置图表类型是折线图
    data: [5, 20, 36, 10, 10, 20]   //设置系列数据
}]
```

1　http://echarts.baidu.com/api.html#echarts。

2　http://echarts.baidu.com/option.html#title。

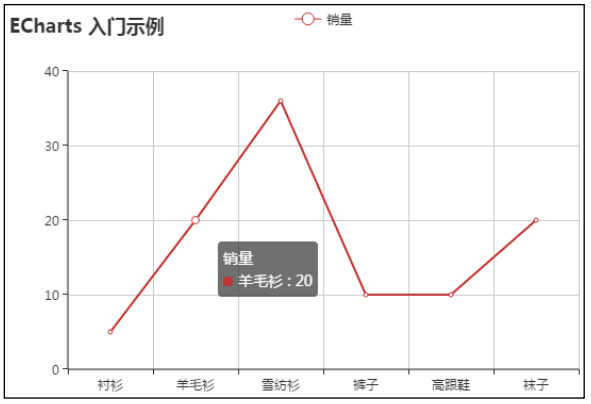


图 6.3　折线图

- **增加图表标签**

为了更好地显示折线图 6 个数据点的标签，在 series 中增加“label(标签)”，效果如图 6.4 所示。代码如下：

```
series: [{                          //设置系列列表
    name: '销量',                   //设置图表系列的名称
    type: 'line',                   //设置图表类型是折线图
     label: {                       //设置显示图表标签
     normal: {
         show: true,
     }
    },
    data: [5, 20, 36, 10, 10, 20]//设置系列数据
}]
```

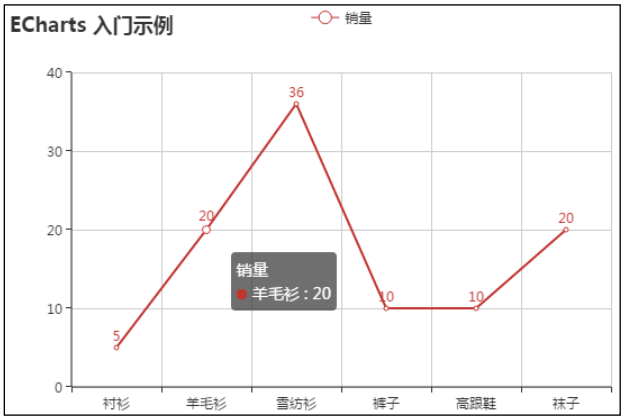


图 6.4　显示标签的折线图

- **增加图表均值和最大值、最小值**

为了更好地理解数据点在均值之上还是在均值之下，可以增加一条“平均值”线，方便读者理解各个数据的大小，效果如图 6.5 所示。代码如下：

```
markLine : {
    data : [{
        type : 'average', name: '平均值'}
            ]
}
```

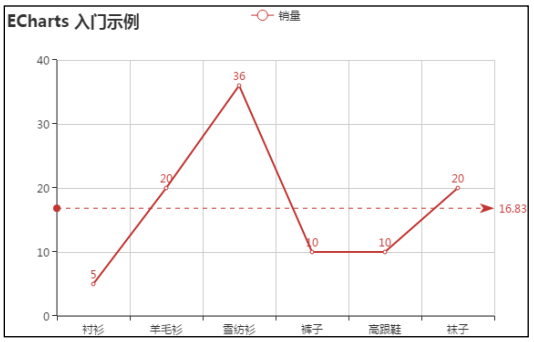


图 6.5　带均值线的图表效果

为了方便读者快速查看数据点中的最大值和最小值，在已绘制的图表中增加相关信息，效果如图 6.6 所示。代码如下：

```
markPoint : {
    data : [
    {type : 'max', name: '最大值'},
    {type : 'min', name: '最小值'}
    ]
}
```

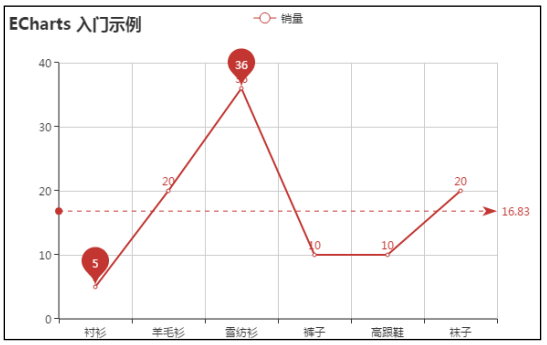


图 6.6　显示最大值和最小值的图表效果

- 主题的使用

主题也称为皮肤，用于修改图表的配色效果，登录官网[1]可查看主题效果并下载合适的主题。为避免制作的图表配色单一，使用“vintage”主题，效果如图 6.7 所示。代码如下：

```
<script src="vintage.js"></script>
var myChart = echarts.init ( document.getElementById ( 'main' ) ,'vint age' ) ;
```

图 6.7　设置图表主题

代码的第一行用于引入“vintage”主题。为方便存储，经常将多个主题放在一个文件夹中。如主题保存在“theme”文件夹，则第一行代码修改如下：

```
<script src="theme/vintage.js"></script>
```

要特别注意本行代码要在引用 ECharts 文件“echarts.min.js”之后，否则主题无法显示，具体代码顺序如下：

```
    <script src="echarts.min.js"></script>
    <script src="theme/vintage.js"></script>
```

代码中 init( )方法的第二个参数用于指定引入的“vintage”主题。

6.1.4　图表中的地图

使用地图前，首先要下载[2]相应的地图，如“北京地图”、“中国地图”或各省地图等。本小节下载的是北京JS版本地图“beijing.js”，并将该地图保存在与“echarts.min.js”文件相同的文件夹中。具体代码如下：

```
<!DOCTYPE html>
<html>

<head>
```

1　http://echarts.baidu.com/download-theme.html。

2　http://echarts.baidu.com/download-map.html。

```
    <meta charset="utf-8">
    <title>ECharts</title>
     <script src="echarts.min.js"></script>
     <script src="beijing.js"></script>
</head>

<body>
    <!-- 为 ECharts 准备一个具备大小（宽高）的 DOM -->
    <div id="main" style="width: 1000px;height:800px;"></div>
    <script type="text/javascript">
     var myChart = echarts.init(document.getElementById('main'));
     var option = {
         title: {
             text: '2014 年北京市各区县人口数（万人）',
             subtext: '数据来源于北京市统计局 http://www.bjstats.gov.cn/',
             left: 'center'
         },
         tooltip: {
             trigger: 'item'
         },
         legend: {
           orient: 'vertical',
           x:'right',
           data:['数据名称']
        },
         dataRange: {
           min: 0,
           max: 250,
           color: ['orange','yellow'],
           text: ['高','低'],          // 文本，默认为数值文本
           calculable : true
        },
         toolbox: {
             show: true,
             orient: 'vertical',
             left: 'right',
             top: 'center',
             feature: {
                 dataView: {readOnly: false},
                 restore: {},
                 saveAsImage: {}
             }
         },
         series: [
```

```
            {
                name: '人口数',
                type: 'map',
                mapType: '北京',//注意，此处地图类型是中文
                roam: false,
                label: {
                    normal: {
                        show: true
                    },
                    emphasis: {
                        show: true
                    }
                },
                data:[
                    {name: '东城区',value: 98.0},
                    {name: '西城区',value: 142.9},
                    {name: '朝阳区',value: 204.2},
                    {name: '丰台区',value: 112.8},
                    {name: '石景山区',value: 38.0},
                    {name: '海淀区',value: 238.5},
                    {name: '门头沟区',value: 24.9},
                    {name: '房山区',value: 79.4},
                    {name: '通州区',value: 70.5},
                    {name: '顺义区',value: 60.9},
                    {name: '大兴区',value: 65.1},
                    {name: '昌平区',value: 58.5},
                    {name: '怀柔区',value: 28.1},
                    {name: '平谷区',value: 40.1},
                    {name: '密云县',value: 43.3},
                    {name: '延庆县',value: 28.2}
                ]
            },

        ]
    };
    myChart.setOption(option);
    </script>
</body>
</html>
```

代码效果如图 6.8 所示。因为 EChats 提供的地图种类有三十余个，所以很多时候用文件夹保存地图，如将文件保存到“map”文件夹，则代码修改如下：

```
    <script src="map/china.js"></script>
```

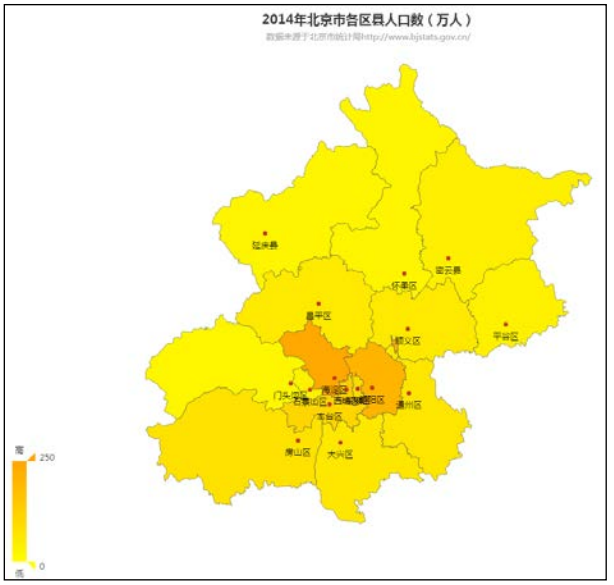


图 6.8　制作图表地图

代码中的“dataRange”用于添加一个数据筛选器，本例中设置了数据范围[0，250]，地图中显示对应的颜色和文本信息。“calculable”用于设置是否启用值域漫游。如图 6.9 所示设置的数据范围是[50，200]，地图中会显示符合条件的区域，不符合条件的区域是白色的。

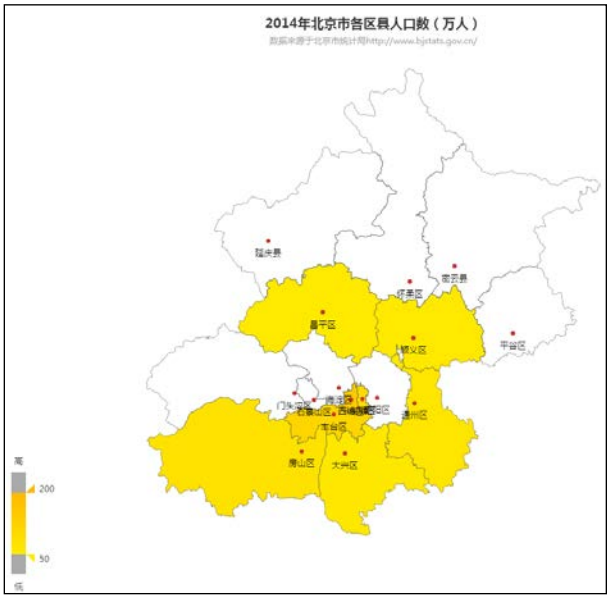


图 6.9　使用图表地图的数据筛选

6.2 标签云

标签云（Tag Cloud），也称文字云，是对关键词的视觉化描述，用于汇总用户生成的标签或一个网站的文字内容。标签一般是独立的词汇，经常按字母顺序排列，通过改变字体大小或颜色来表现其重要程度，所以标签云可以灵活地依照字母顺序或热门程度来检索一个标签。大多数标签本身就是超级链接，直接指向与标签相关的一系列条目 [1]。照片共享网站Flickr是最先使用标签云的高知名度网站，随后标签云逐渐被各大网站使用并流行起来。

标签云是一种数据可视化手段，目前在国内外的数据新闻中经常使用，通过标签文本的字号大小反映其权重，如用字号表现标签出现的频度，频度越高字号越大。

标签云工具很多，其属性一般有四种，“字号”属性一般与数量相关，数量越多字号越大。“排列”属性用于设置文本排列方式，“颜色”属性用于固定渐进色、是否增加背景等，“字体”属性用于设置文本字体。

早期使用 Excel 等软件制作简单的标签云，方法复杂，效果欠佳。随后，开始使用 DIV、CSS 和 JS 等技术制作复杂标签云，效果好但需要编写代码，制作周期长。现在，大量标签云制作工具软件如雨后春笋般出现并被大多数没有代码编写经验的大众接受并使用。

常用的标签云制作工具有 Wordle、TagCloud、Imagechef 、WordItOut、Tagcrowd、Tagul 和 Tagxedo 等，具体功能如下。

- Wordle[2]工具支持三种内容来源，自己输入、URL和del.icio.us的账号。可以设置样式、字体、布局和配色方案等，但是并不支持导出为图片，可以打印、与其他Wordle用户分享，还提供可以嵌入网页的HTML代码。在Wordle的高级设置中，实现了一些颜色和权重的设置，但不支持中文。
- TagCloud[3]是 3D文字云，可以保存为Flash格式，但不支持中文。
- Imagechef[4]可视化效果较弱，主要是轮廓运用，不计算文本的频次，只是将这些文字无规律地重复填充在设置的轮廓中，支持中文和轮廓自定义。
- WordItOut[5]支持中文，可以设置权重和颜色，还可以选择颜色变化规则。
- Tagcrowd[6]不支持中文，而且单词少，但是可以生成PDF。
- Tagul[7]功能强大，需要申请账号。支持中文，颜色、形状、字体均可以设置。

1 http://zh.wikipedia.org/zh-cn/%E6%A0%87%E7%AD%BE%E4%BA%91。
2 http://www.wordle.net。
3 http://tagcrowd.com/。
4 http://www.imagechef.com。
5 http://worditout.com/。
6 http://tagcrowd.com/。
7 http://www.tagul.com。

- Tagxedo[1]是一个功能强大的文字云工具，不仅支持中文而且功能强大。

制作中文数据新闻时经常使用的标签云工具是 Tagul 和 Tagxedo。因为这些工具的服务器均在国外，所以有些时候速度不佳，建议下午使用。

6.2.1 标签云制作工具 Tagul

登录 Tagul 官方网站 http://www.tagul.com，注册账号并登录。账号信息显示主界面右上角，如图 6.10 所示。主界面显示已经制作完成的标签云，如图 6.10 所示包含了 3 个制作完成的标签云，分别显示了标签云的名称、创建日期和是否公开等信息。注意，3 个标签云的“Visibility”不同，既有公开的(“Public”)，也有私有的(“Private”)，如果希望标签云分享到社交网站，标签云必须是公开的。单击“Delete”按钮可以删除所选标签云，“Duplicate”按钮可以复制标签云进行备份。若制作完成的标签云过多，可以使用搜索窗口筛选标签云，也可以单击页面右下角的“Previous”和“Next”按钮翻页查找。

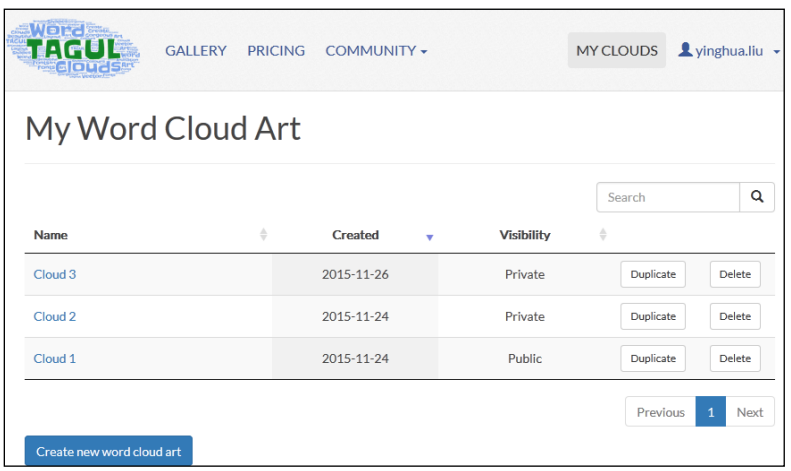


图 6.10　Tagul 主界面

创建标签云。单击如图 6.10 所示左下角的“Create new word cloud art”按钮可以创建一个新的标签云。在打开的窗口的“Name”文本框中输入标签云的名称，如“Tools”，如图 6.11 所示。不建议为标签云设置中文名称。

编辑标签云文字。在如图 6.11 所示的“Words”选项卡中输入文字。文字可以通过“Import words”按钮从某个 URL 导入，也可以导入 Excel、CSV 或纯文本格式的信息。

可以单击“Add”或“Remove”按钮添加或删除文字。大部分情况下标签云中的文字并不是特别复杂，自行输入即可。“Clear all” 按钮用于删除所有文本。排在前面的文字显示得大，排在后面的文字显示得小。

1　http://www.tagxedo.com/。

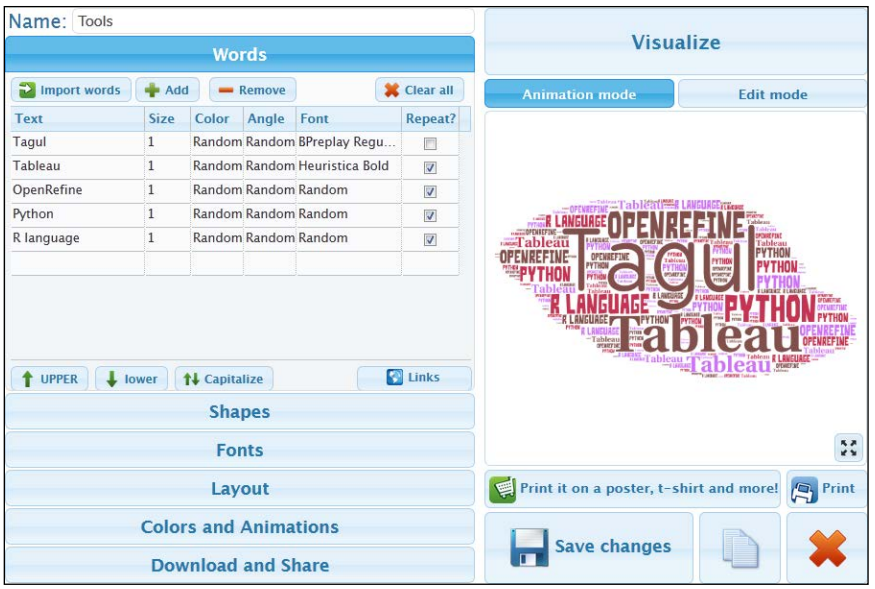


图 6.11　创建标签云

文字的大小、颜色、角度和字体默认为“Random”，即文字这 4 个属性均为随机的，可以根据个人喜好修改属性，本例中修改了前两个文字的字体。每次修改后单击右上角的“Visualize”按钮可以查看效果。

“Repeat”选项决定该文字是否重复，本例中文本“Tagul”没有勾选该选项，在右侧的效果中可以看到该文本仅出现了一次，而其他文本均出现多次。

“UPPER”选项将文本均变成大写，“lower”选项将文本均变成小写，“Capitalize”选项将文本首字母变成大写。

保存标签云。编辑标签云后要及时保存，单击右下角的“Save changes”按钮可以保存修改。

设置标签云形状。如图 6.12 所示的“Shapes”选项卡提供了常用的标签云格式，可以根据需要为标签云选择合适的形状。也可以添加个性化形状，单击左上角的“CLICK HERE TO ADD YOUR IMAGE”按钮可以添加本地计算机的图片或者图片的 URL 作为标签云形状。

设置标签云字体。如图 6.13 所示的“Fonts”选项卡可以设置标签云的文本字体，默认情况下，Tagul 提供的均是英文字体。可以单击上面的“Click here to add your font”按钮添加本地计算机的字体，以更好地呈现文字效果。若标签云中包含中文字体，将显示乱码或者不显示该中文，建议添加中文字体以正确显示，如图 6.13 所示添加了中文古隶简“FZGuLi-S12S”字体。

设置标签云布局。如图 6.14 所示的“Layout”选项卡可以设置标签云的文本排列位置，以及文字的数量和大小。

设置标签云颜色和动画效果。如图 6.15 所示的“Colors and Animations”选项卡可以设置标签云的文本颜色、动画速度、动画是否放大、是否旋转、背景颜色、翻滚框文字颜色和翻滚框颜色。

图 6.12 “shapes”选项卡

图 6.13 “Fonts”选项卡

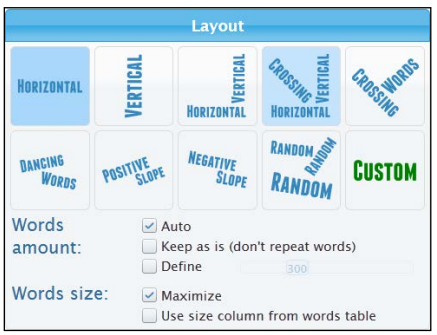


图 6.14 “Layout”选项卡

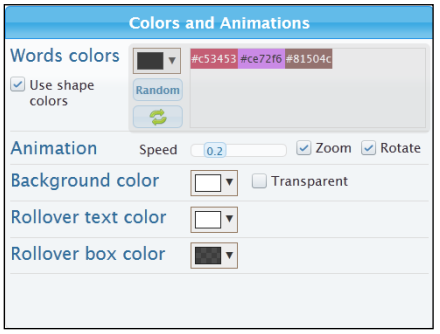


图 6.15 “Colors and Animations”选项卡

设置标签云模式。在“Animation mode”(动画模式)下，当用鼠标点选文本时，呈现动画效果，如图 6.16 所示。在“Edit mode”(编辑模式)下，可以编辑选择的文本，如文本大小、颜色和位置等，如图 6.17 所示。

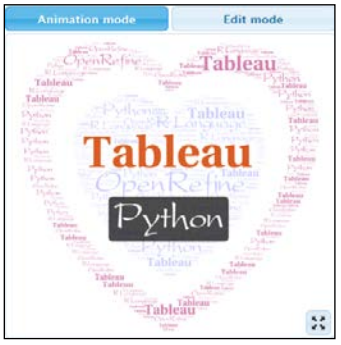


图 6.16 动画模式

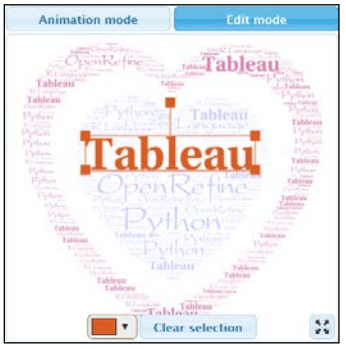


图 6.17 编辑模式

查看标签云效果。单击主界面右上角的“Visualize”按钮可以查看标签云效果，特别适合展示有动画效果的标签云。

下载和分享标签云。打开“Download and Share”选项卡可以下载标签云，如图 6.18 所示。下载的标签云可以保存为 PNG 格式的位图图片，也可以是 EPS 或 SVG 格式的向量图片。如图 6.19 所示是一个中文标签云下载的 PNG 格式效果。

图 6.18　下载或分享

图 6.19　中文字体效果

在“Download and Share”选项卡中还可以分享标签云，如设置标签云是否公开，是否分享到 Google+、Facebook、Twitter 和 Link，或者将标签云发送邮件给他人，或者下载嵌入 Web 页面的代码，将该链接嵌入到某个网页中，如图 6.20 所示。注意，标签云设置为私有时不能嵌入到其他 Web 页面。

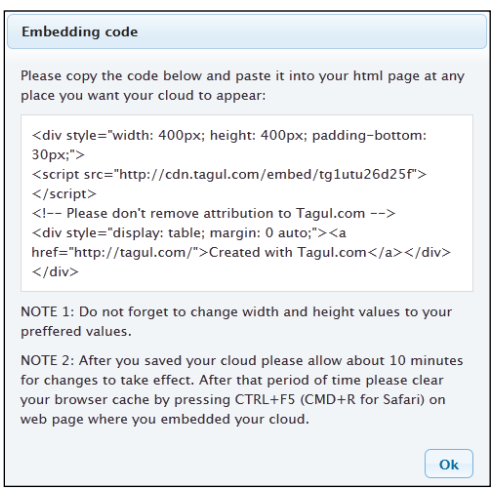


图 6.20　嵌入 Web 页面的代码

6.2.2　标签云制作工具 Tagxedo

下载并安装Microsoft Silverlight[1]，这是使用Tagxedo制作标签云的必备插件。

1　https://www.microsoft.com/silverlight/。

Microsoft Silverlight 是一个跨浏览器、跨平台的插件，为网络带来下一代基于.NET framework 的媒体体验和丰富的交互式应用程序。Silverlight 提供灵活的编程模型，并可以很方便地集成到现有的网络应用程序中。Silverlight 可以对运行在 Mac 或 Windows 系统上的主流浏览器提供高质量视频信息的快速低成本传递。

登录 Tagxedo 官方网站 http://www.tagxedo.com/，主界面如图 6.21 所示。左侧是功能选项，右侧是标签云展示区，右下角的“FullScreen”和“Zoom”按钮用于设置标签云的呈现效果。“FullScreen”按钮用于设置是否全屏显示。“Zoom”按钮用于设置标签云放大或缩小的比例，范围是 50%～2000%，单击“Fit”按钮标签云会适应窗口大小，单击“In”按钮标签云会增大比例，单击“Out”按钮标签云会减小比例。

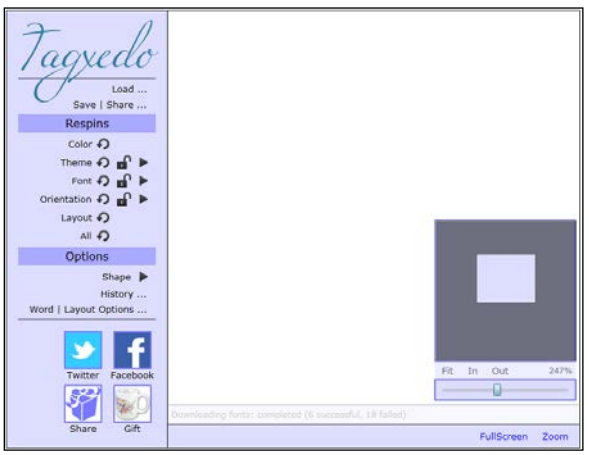


图 6.21　Tagxedo 主界面

导入标签云文字。单击左侧的“Load”选项，在打开的“Load Menu”对话框中输入文字，如图 6.22 所示。

图 6.22　“Load Menu”对话框

文字的输入方式有三种。第一种方式，单击“Browse”按钮打开某个文件后单击“Submit”按钮，将文件中的文字导入，注意，Tagxedo 只能导入文本格式的文件。也可以在“Webpage”中输入某个 URL 地址后单击“Submit”按钮，Tagxedo 将导入此 URL 地址中的文字。第二种方式，在“Enter Text”文本框中直接输入或复制文字后单击“Submit”按钮。第三种方式，运行 XAP 文件，这是 Silverlight 应用程序编译后的一种文件格式，包括 Silverlight 应用程序所需的全部文件，如程序集、资源文件等。注意，Tagxedo 不能将三种输入的内容共同显示，即只显示某一种输入方法包含的文字。在“Enter Text”文本框中可以使用【Ctrl】+【A】、【Ctrl】+【C】和【Ctrl】+【V】快捷键。

设置标签云的主题颜色。单击“Theme”右侧的三角形按钮，在打开的“Theme Menu”面板中可以设置文字的主题颜色，如图 6.23 所示。也可以单击底部的“Add Themes”按钮增加个性化主题。单击“Theme”右侧的锁头型符号可以锁定或解锁该选项。

设置标签云字体。单击“Font”右侧的三角形按钮，在打开的“Font Menu”面板中可以设置文字的字体，如图 6.24 所示。既可以单击“Use Local Fonts”按钮使用本地字体，也可以单击底部的“Add Fonts”按钮增加个性化字体。注意，默认字体均是英文字体，若有中文或其他非英文文字，可以单击右下角的“Add Fonts”按钮添加字体，上传字体后即可使用。

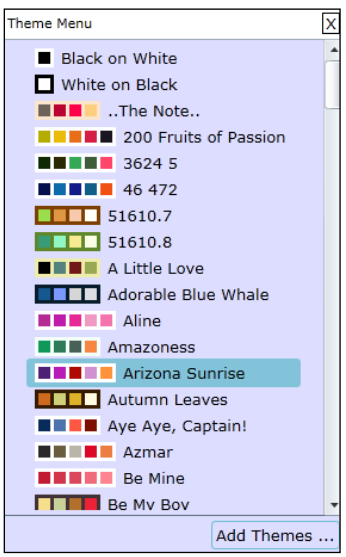


图 6.23　设置标签云的主题颜色

图 6.24　设置标签云字体

设置标签云方向。单击 “Orientation”右侧的三角形按钮，在打开的“Orientation Menu”面板中可以设置文字的方向，如图 6.25 所示。方向包括“Any”（任意方向）、“Horizontal”（水平方向）、“Vertical”（垂直方向）和“H/V（Orthogonal）”（水平垂直共存方向）四种。

设置标签云形状。单击“Shape”右侧的三角形按钮，在打开的“Shape Menu”面板中可以设置文字的形状，如图 6.26 所示。

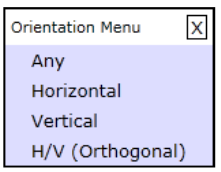


图 6.25　设置标签云方向

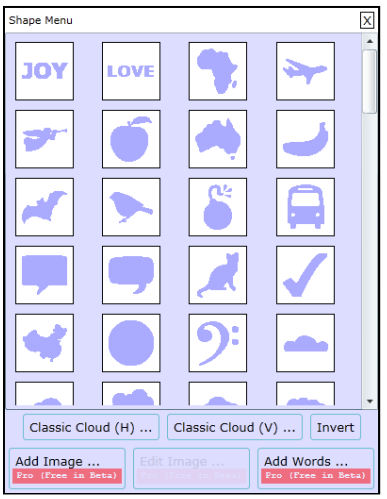


图 6.26　设置标签云形状

使用标签云历史。单击“History”选项将展现操作历史，如图 6.27 所示。

设置标签云文字和布局。单击 “Word|Layout Options”选项，在打开的“Option Menu”面板中（包含四个选项卡）可以设置标签云的布局。如在“word”选项卡中单击“Apply NonLatin Heuristics”后面的“No”按钮，再单击“Accept”按钮应用该操作，可以保证输入的文字词组不分散。在该面板中可以对文字、样式和颜色等重新刷新或者组合。

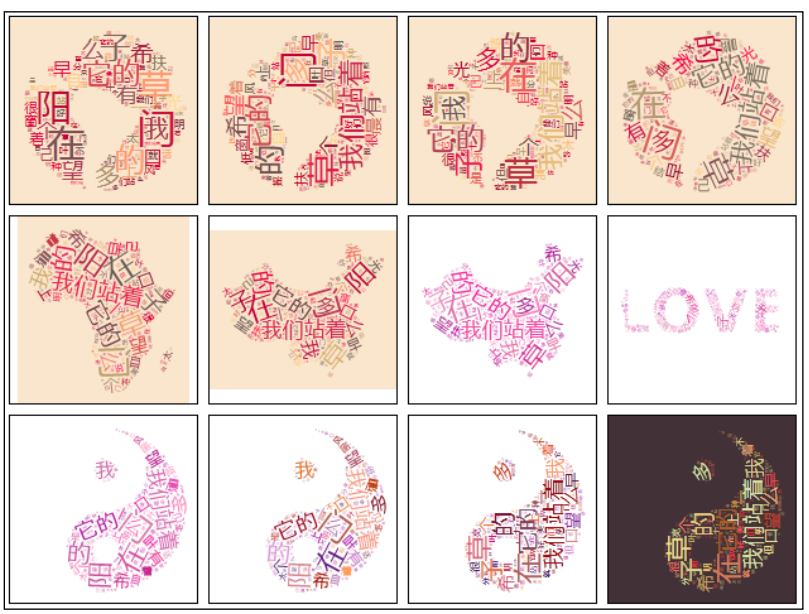


图 6.27　展现标签云历史

保存标签云。单击“Save|Share”选项，在打开的“Save Menu”面板中可以选择“Image”选项卡将标签云保存为图片，也可以在“Web”选项卡中将标签云保存到网络上，如放到个人博客中。还可以在“Print”选项卡中打印标签云，在“Advanced”选项卡中进行高级设置，如图 6.28 所示。注意，保存的文件越大图片越清晰。

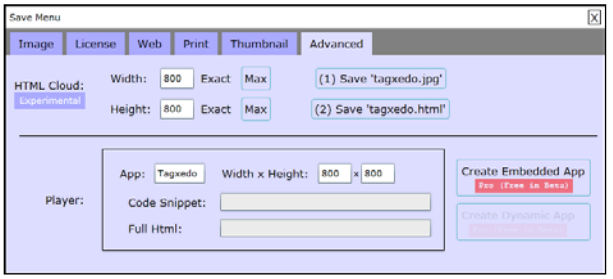


图 6.28　保存标签云

分享标签云。可以将标签云分享到 Twitter 和 Facebook 等，甚至还可以将标签云图案印刷在 T 恤、被单或鼠标垫上，如图 6.29 所示。

图 6.29　标签云印刷在杯子上

6.3 关系图制作工具 PeoplePlotr

获取的数据如果存在上下级关系，可以使用类似族谱、组织结构图和层次关系图的方式进行可视化呈现，如在可视化社交媒体信息时展现人物之间的关系。

PeoplePlotr 是一个基于网页的应用，由英国伦敦的 Webalon 团队开发，该工具可以实现上述功

能。如图 6.30 所示是使用 PeoplePlotr 制作的一个族谱关系图。

图 6.30　使用 PeoplePlotr 制作的族谱关系图

使用 PeoplePlotr 制作的人物关系图是动态的，包含文字、图片和视频等多媒体信息，可以快速修改人物关系图的形状，设置每个关系图节点的大小、背景图、透明度和颜色等。还可以为每个关系图节点添加说明信息、图片、音频和视频等。查看某个节点的详细信息效果如图 6.31 所示，在其左上角可以选择查看该节点的多媒体信息。

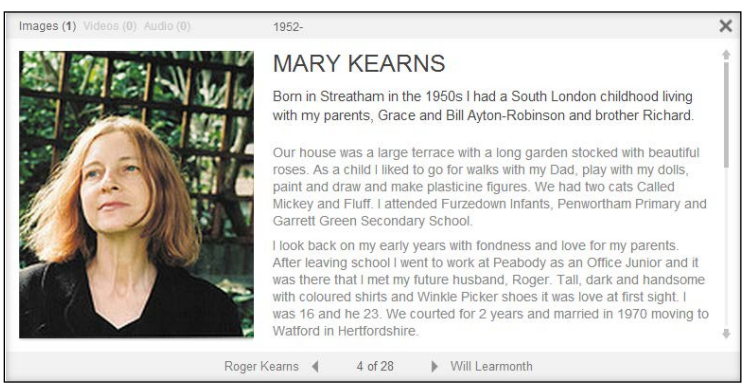


图 6.31　查看某个节点的详细信息

如果关系图包含的节点过多，可以使用如图 6.32 所示的搜索功能筛选节点，也可以使用时间线按时间顺序显示节点（如图 6.33 所示），或者通过节点浏览筛选。

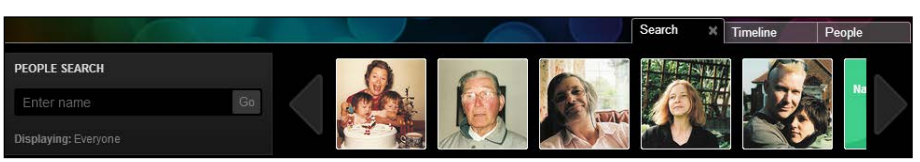


图 6.32　查找功能

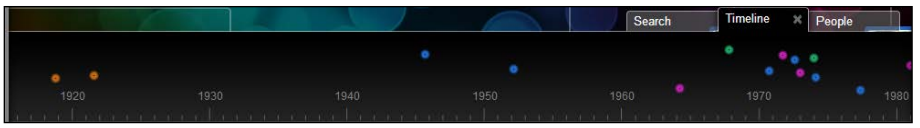


图 6.33 时间线显示功能

PeoplePlotr 简单易学，为节点添加内容非常简单。完成后的作品可以分享给家人或好友，也可以嵌入到网站或博客上（该功能需要收费）。PeoplePlotr 的缺点是不支持直接从本地上传图片、音频和视频，所有内容必须通过网络链接或使用 Flickr 账户添加，这意味着用户使用本地数据必须上传到网络后才能使用，有可能产生隐私泄露的问题。

创建账号。登录官方网站 [1]，创建一个免费账号。免费账号只允许创建一个关系图，最多包含 30 个节点，可以添加网络或Flickr图片，还可以免费设置背景图片并与他人分享作品，但不能将作品嵌入网站。在官方网站页面的下方创建账号，如图 6.34 所示，输入用户名、电子邮件地址和密码，同意PeoplePlotr的条款和条件，单击“Sign up for free account”按钮即可完成账号的创建并自动登录账号。PeoplePlotr不使用邮件确认等方法开通账号，所以账号创建后可以立即使用。

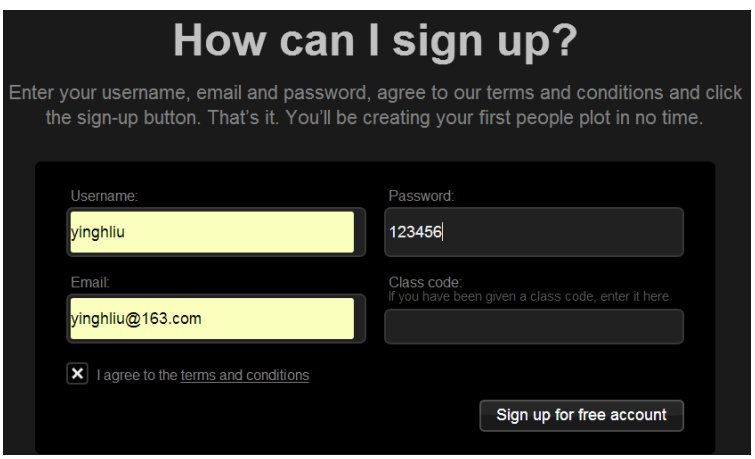


图 6.34　创建一个免费账号

登录账号。账号创建后可以自动登录使用。也可以在 PeoplePlotr 官网页面的“LOG IN”处登录账号，注意在登录账号时，密码是明文显示的，如图 6.34 所示。

新建关系图。填写主题和描述，设置背景图等。下面以经典电视剧《琅邪榜》为例，创建剧中众多角色关系图。新建关系图时，只有主题是必须填写的，其他内容可以后期编辑，单击“CREATE NEW PLOT”按钮新建关系图，如图 6.35 所示。由于 PeoplePlotr 服务器在国外，所以保存关系图时需要一些时间。

1　http://www.peopleplotr.com/。

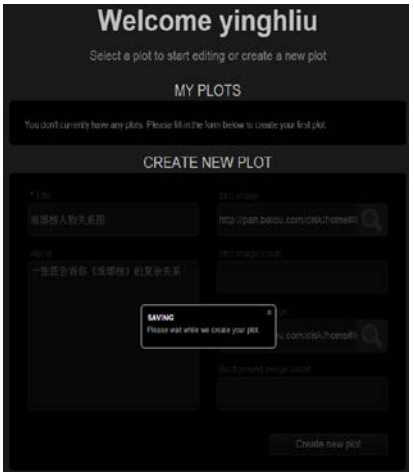


图 6.35　创建一个关系图

设置关系图编辑模式。默认关系图的效果如图 6.36 所示，包含 6 个节点，两级关系。其中，一级包含 2 个节点，二级包含 4 个节点。

图 6.36　默认关系图

关系图左下角的“Edit mode”选项可以设置编辑模式，如图 6.37 所示。单击“On”按钮，进入可编辑模式，管理面板的右侧将出现说明，如图 6.38 所示，可以为关系图和节点添加、修改信息。单击“Off”按钮，关闭编辑模式，进入浏览模式，只能查看关系图的整体效果和每个节点的详细信息。

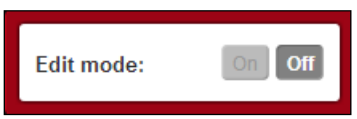


图 6.37　设置编辑模式

编辑关系图的基本信息。进入编辑模式，单击非节点处，进入关系图的基本信息编辑状态，如图 6.38 所示。在此状态下可以修改关系图的主题、描述和设置背景图等。注意，关系图的背景图只

能来自于网络，可以通过浏览器搜索合适的图片，本案例使用的背景图片来源于“http://uploads.rayli.com.cn/2015/1019/1445245677931.jpg”，可调整透明度等以达到合适的图片显示效果。单击“Save”按钮保存修改，单击“Revert”按钮进行恢复。

编辑节点的基本信息。进入编辑模式，单击某个节点，进入节点的基本信息编辑状态。本案例选择的节点是“靖王（萧景琰）”，如图 6.39 所示，在此处可以修改该节点的名称、出生日期、死亡日期、性别、描述、图片来源、图片大小、分类、图片箭头方向和文本等。本案例“靖王（萧景琰）”节点的图片来源于“http://c.hiphotos.baidu.com/baike/c0%3Dbaike116%2C5%2C5%2C116%2C38/sign=ba6d47a20af3d7ca18fb37249376d56c/d439b6003af33a87cb1f0086c05c10385343b53f.jpg”。

图 6.38　编辑关系图

图 6.39　编辑节点

划分类别。进入编辑模式，编辑关系图包含的类别，类别用于划分关系图中的节点，本案例将电视剧《琅琊榜》的所有角色分为四类，分别是“其他派”、“太子派”、“誉王派”和“靖王派”，如图 6.40 所示，并为四种类别设置不同的显示颜色。划分类别后，在浏览模式中查看关系图时，可以快速地通过颜色确定节点类别。单击“Save”按钮保存修改。

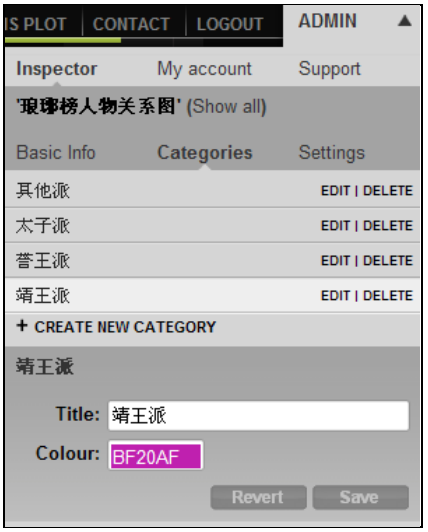


图 6.40　划分类别

添加节点。进入编辑模式，单击左下方的“Male”或“Female”按钮并拖动到关系图中合适的位置，效果如图 6.41 所示。

图 6.41　添加节点

删除节点。进入编辑模式，选择要删除的节点，在“Basic info”选项卡中单击“Delete”按钮即可，如图 6.39 所示。注意，删除的节点无法恢复。

添加连接线、文本和图例。单击图 6.41 中的“Basic Tree”，添加节点间的树形关系，“Vertical Line”和“Horiz Line”分别用于添加垂直关系线和水平关系线，“Corners”用于添加四种拐角线，“Text”用于添加文本，主要说明节点之间的关系。“Legend”用于为关系图添加图例，如图 6.42 所示。图例显示了四种类别的颜色。

图 6.42　添加图例

移动和调整节点。移动节点时，选择一个节点并将其拖动到合适的位置，也可以使用上下左右箭头微调节点的位置。

为节点添加多媒体信息。选择一个节点，在“Images/Videos”中单击“ADD NEW MEDIA”按钮可以为节点添加多媒体信息，如图片、音频和视频。需要注意的是，多媒体信息必须是网络信息，不能直接使用本地计算机保存的信息，如图 6.43 所示。

图 6.43　为节点添加多媒体信息

浏览作品。本案例部分效果如图 6.44 所示，可以根据个人爱好添加更多的节点，设置关系图的个性化布局。如果关系图节点多，屏幕无法全部显示，可以按住鼠标左键左右拖动查看，或者单击屏幕中的上、下、左、右三角形图标查看。单击某个节点，可以浏览该节点详细信息。如图 6.45 所示是“越贵妃（贤妃）”节点的浏览效果。

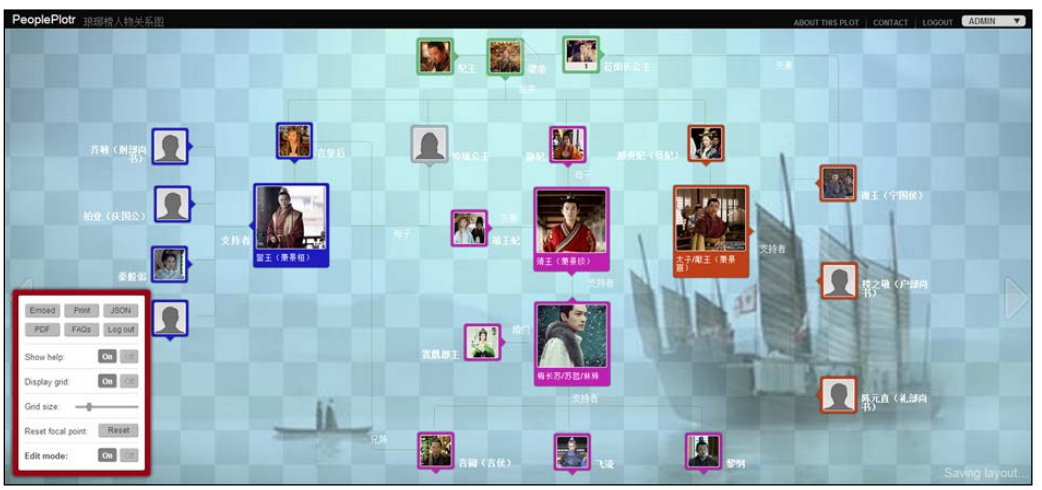


图 6.44　案例部分效果

图 6.45　浏览“越贵妃（贤妃）”节点信息

编辑账号。“My account”选项卡显示当前的账号信息，如账号名称、E-mail 地址和账号等级等，如图 6.46 所示。单击“Change”按钮可以修改账号密码，单击“Upgrade”按钮可以更改账号等级，级别越高功能越强，但价格也越高。

导出作品。在编辑模式下，可以单击“Embed”按钮得到作品 URL 后嵌入到其他 Web 页面，也可以单击“PDF”按钮将作品输出为 PDF 格式的文件，如图 6.47 所示。

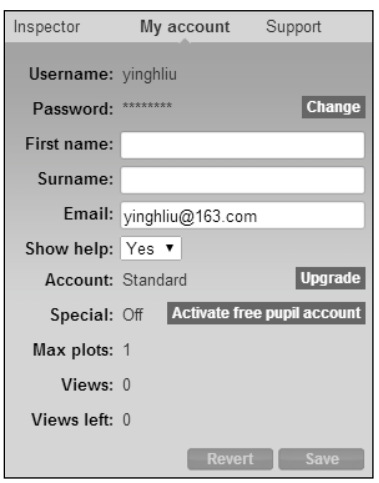


图 6.46　编辑账号

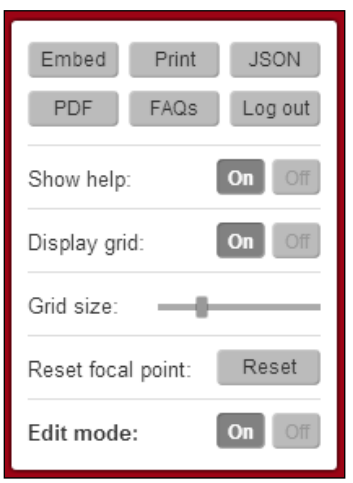


图 6.47　导出作品

注意，导出 PDF 文件前需要下载并安装 Webshot（网址是 http://www.websitescreenshots.com/）。单击“PDF”按钮后会给出相应的提示信息。

6.4 语义万维网服务 Open Calais

Open Calais 网络服务是路透社新近推出的语义万维网服务（Semantic Web），是基于 Web 2.0 的服务。该服务方便信息发布者自动在其内容中为人物、地点和事件等设定元标签（metatag）。元标签是“关于数据的数据”。例如，一本书的标题和作者是关于书的元数据。在信息系统中，标签是分配给一块信息的非层次关键字或术语（如网络书签、数字图像或计算机文件）。这种元数据有助于描述一个项目，并允许它再次通过浏览或搜索找到。

Open Calais 使用自然语言处理（NLP）和数以百计的机器学习算法，通过路透社编辑团队数年的经验，为业界提供最好的特征提取和相关性组合。对普通用户来说，这个过程也非常简单。用户提供非结构化文本到提取引擎（如新闻或博客等）检测和定位文本。Open Calais 允许用户每天最多上传 5000 个文档。

Open Calais 用户范围广泛，既可以是博主，也可以是投资银行家，用户使用 Open Calais 的目的是相同的，都是想从巨大的数据中发现、提取或获得期待的知识。Open Calais 可以描述许多不同种类的信息。例如，元标签服务中元标签可以说明以下内容。

实体（Entities），如公司、人物、地点或产品等。

关系（Relationships），如张三在某公司工作。

事实（Facts），如张三是一位 51 岁的男性 CEO。

事件（Events），如张三被任命为某公司的董事会成员。

主题（Topics），如故事讲述制药行业并购案。

Open Calais 网络服务的地址是 http://www.opencalais.com/opencalais-demo/。

打开 Open Calais 后处理从文本中提取的信息，以 RDF 格式返回语义元数据。Open Calais 的优点是有上下文导航，可以精确定位相关的公司、人物和行业；更有针对性的消息，如获取公司和有关行业高度相关的、有针对性的新闻；可以快速地发现并返回隐藏在文本中的相关事实和事件。

对中国用户来说，Open Calais 最大的缺点是不支持中文。本案例使用的内容来自“http://www.usatoday.com/story/news/politics/elections/2016/04/26/tuesday-primaries-could-make-nomination-unreachable-sanders/83561132/”。

首先浏览待分析的 Web 页面。此页面是 2016 年美国选举的相关报道，主要内容是关于希拉里·克林顿（Hillary Clinton）和参议员伯尼·桑德斯（Bernie Sanders）的选举进展情况，具体内容如图 6.48 所示。

复制该 Web 页面。按住鼠标左键选择需要处理的文本部分，虽然 Open Calais 仅适合分析文本，但也可以处理照片出处等信息，所以多媒体信息虽然未能详细处理，但对文本分析有帮助，使用【Ctrl】+【A】快捷键选择全部内容，再使用【Ctrl】+【C】快捷键进行复制。

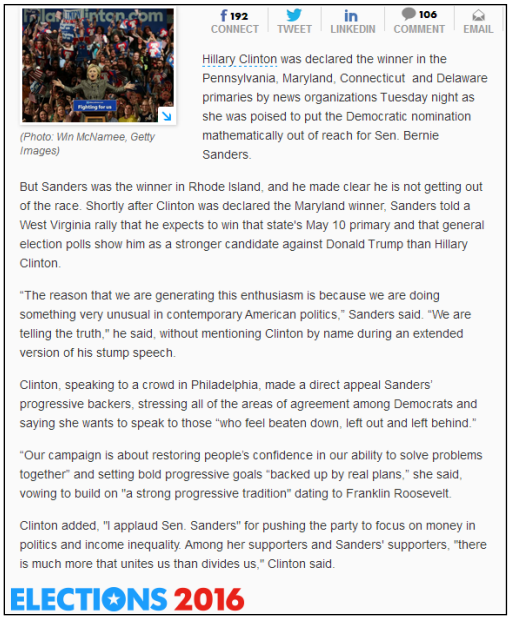


图 6.48　待处理的 Web 页面

使用 Open Calais 分析文本。进入 http://www.opencalais.com/opencalais-demo/，打开 Open Calais 服务。使用快捷键【Ctrl】+【V】粘贴复制的内容，如图 6.49 所示，然后单击“TAG IT”按钮。用户也可以拖放或上传一个 PDF/XML 格式文件获取元标签。

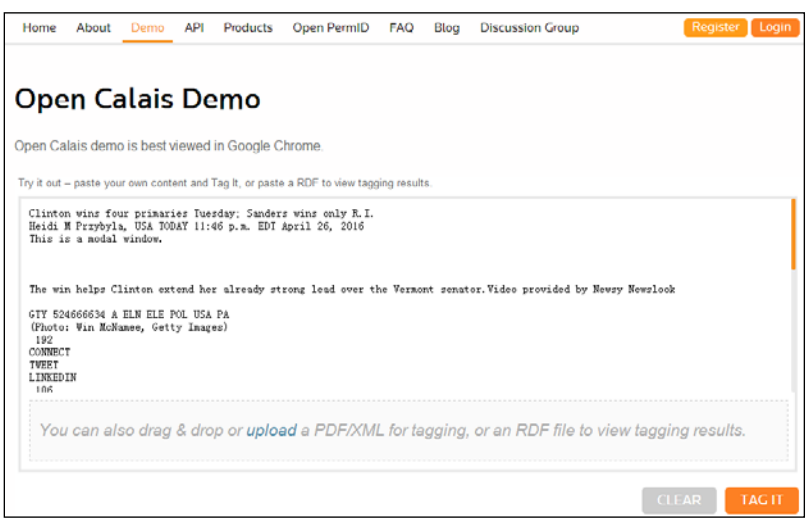


图 6.49　Open Calais 服务界面

查看分析效果。Open Calais 处理后的效果如图 6.50 所示。

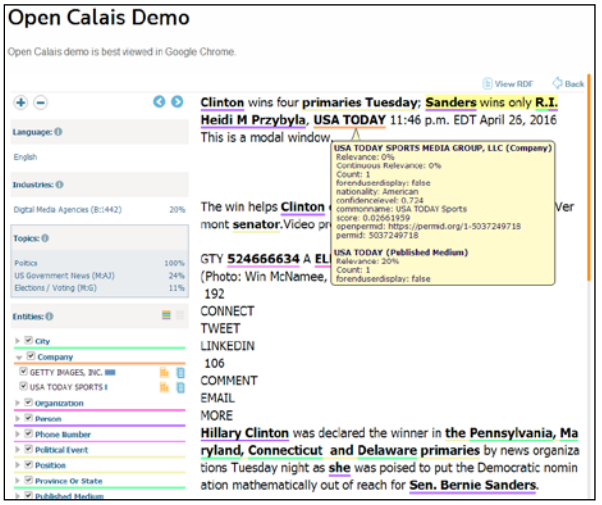


图 6.50　Open Calais 处理结果

如图 6.50 所示处理结果的左侧显示了文档的语言、数字媒体机构和主题等信息，该文档的“政治”主题占 100%，“美国政府新闻”占 24%，“选举/投票”占 11%，还显示了实体、事件和社会标签等信息，如图 6.51 和图 6.52 所示。

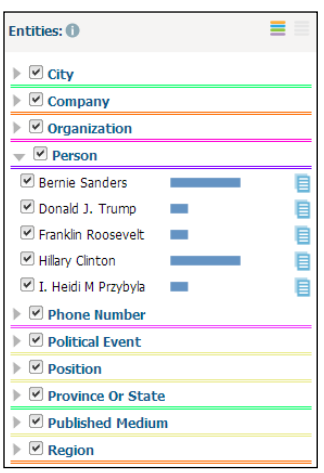


图 6.51　实体信息

图 6.52　事件和社会标签信息

在图 6. 51 中，单击三角形标志可以展开实体的详细信息。图 6.51 中包含五位个人信息，将鼠标指针移动到某个人的名字上，可以详细显示此人信息，如人物关联度、人物出现的次数、国籍和置信水平等信息。如图 6.53 所示为显示人物“Bernie Sanders”的具体信息。社会化标签（Social Tags）

尝试模拟一个人会如何标记一个特定内容，这是通过比较文章的全文而得到的。如图 6.54 所示为显示相应的版本信息。

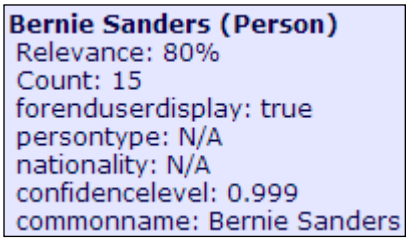


图 6.53　人物“Bernie Sanders”详细信息

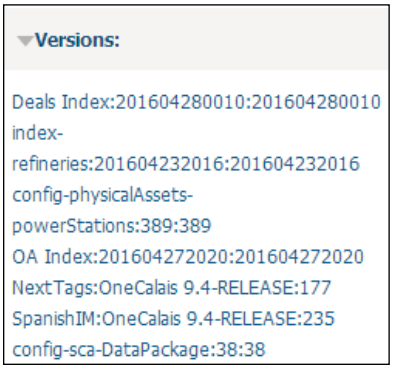


图 6.54　版本信息

如图 6.50 所示处理结果的右侧按实体颜色高亮显示了文档的内容，方便用户查看。用户还可以单击“View RDF”按钮查看元标签的结果，如图 6.55 所示。注意，XML 文件不显示与之关联的任何样式信息，文档以树型结构显示。

Open Calais 帮助数据新闻制作者快速了解文章内容，发现实体及实体之间的关系，掌握实体的背景资料，清楚多媒体信息来源，发掘活动和事实。

Open Calais 的相关信息可以查询 http://www.opencalais.com/about-open-calais/，常见问题可以在 http://www.opencalais.com/open-calais-faq/中得到解答。

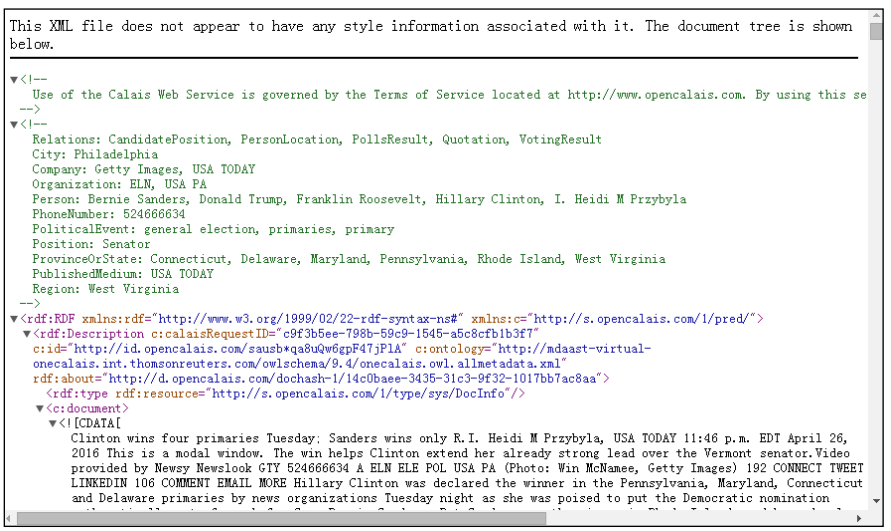


图 6.55　查看元标签的结果

6.5 HTML5 网站制作模板

HTML5 简称 H5，是超文本标记语言（HTML，Hypertext Markup Language）的第五次重大修改标准。2014 年 10 月 29 日，万维网联盟宣布该版本标准规范制定完成。

Web 网页是由超文本标记语言实现的，即通过整合使用的其他 Web 技术（如脚本语言等）创造的功能复杂的页面。早期的 HTML5 在不同的浏览器上呈现的效果差异很大，现在由于手机等移动端网络设备的兴起，且手机浏览器相对统一，解决网页兼容性难度大大降低。微信的流行也带动了 HTML5 作品的发展和分享。

例如，财新实验室制作的动态数据新闻，主要工具包含 HTML5、CSS3 和 JavaScript 等，其中 HTML5 用于绘制图形和动画，CSS3 实现排版，JavaScript 用于处理交互和动画。

应用于移动平台（如手机、平板电脑等）的HTML5 微场景制作工具很多，使用用户较多且功能丰富的国内工具主要有易企秀 [1]、微页 [2]、Epub360 意派 [3]、MAKA[4]、有赞（口袋通）[5]、快站（搜狐）[6]等。这些工具帮助用户快速制作和编辑移动平台Web页面，并分享到社交网络，也可以收集潜在客户或其他反馈信息等。

HTML5 微场景制作与运用流程包含四个步骤，首先是素材准备，其次是场景制作，第三是发布宣传，最后是收集用户信息。其中，素材准备需要的时间较多，要根据作品的创意准备可以烘托氛围的音乐、视觉效果好的图片和有感染力的文字等。

应用于PC或移动平台的HTML5 网站（网页）制作工具也很多，其中国外的Wix[7]和国内的起飞页 [8]是基于HTML5 技术拖拽即可视的操作方式中的佼佼者，这种制作网站（网页）的方法让任何不会编写代码的普通用户也可以设计并创建个性化的HTML5 网站。

本节以 Wix 为例简要说明如何制作 HTML5 网站。使用 Wix 需要注册免费账号。单击 Wix 官网右上角的“Sign In”按钮，如有账号则登录，若没有账号则进入申请账号页面，如图 6.56 所示。

选择模板。使用账号登录后，在左侧的“Categories”区域中选择合适的模板类型，如选择“Weddings & Celebrations”，在页面右侧会显示该类型的模板效果，如图 6.57 所示。

1　易企秀官网 http://eqxiu.com。
2　微页官网 http://www.weiye.me。
3　Epub360 意派官网 http://www.epub360.com/。
4　MAKA 官网 http://www.maka.im/。
5　有赞（口袋通）官网 http://www.youzan.com/。
6　快站（搜狐）官网 http://zhan.sohu.com/。
7　Wix 官网 http://www.wix.com。
8　起飞页官网 http://www.qifeiye.com/。

图 6.56　Wix 注册免费账号页面

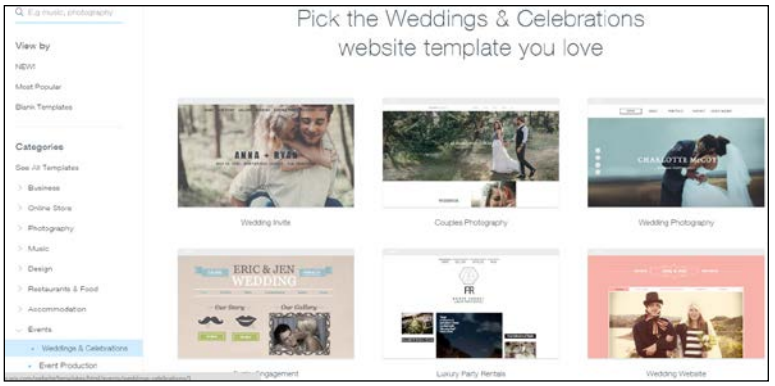


图 6.57　查看模板效果

预览模板效果。将鼠标指针移动到选中的模板上，会显示模板的价格，如图 6.58 所示，单击“Info”选项可以查看该模板的详细信息，如适合展示哪类信息等，如图 6.59 所示。

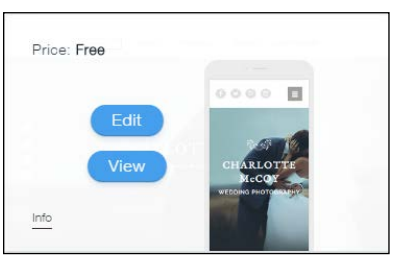


图 6.58　查看模板价格

图 6.59　模板详细信息

单击“View”按钮浏览网站效果。默认情况下显示 PC 端（desktop view）效果，如图 6.60 所示。单击左上角的“Mobile view”可以显示移动端效果，如图 6.61 所示。可以用鼠标向下拖曳滚动条或

通过导航栏查看完整的网页效果。

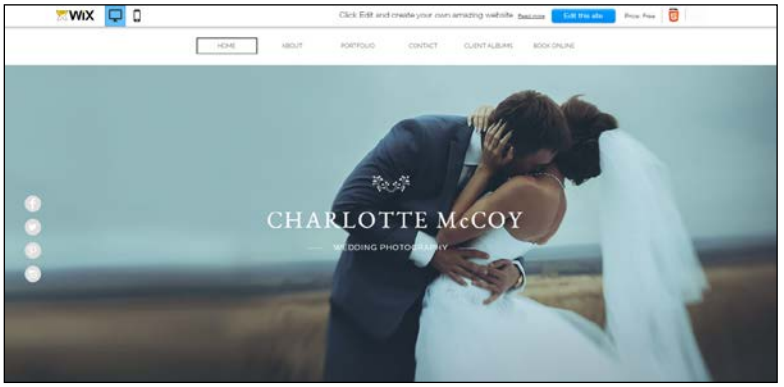


图 6.60　PC 端效果

编辑模板。单击“Edit this site”按钮进入模板编辑状态，编辑按钮如图 6.62 所示。“Background”按钮用于使用预制多媒体设置背景的颜色、图片或视频。

使用“Add”按钮可以添加文字、图片、图集、按钮、图形、视频、音乐、社会媒体、联系人、菜单、列表和博客等对象。

以添加文字为例。单击“Add”按钮，在打开的菜单中选择“Text”，并选择某种字体，直接拖动到网页上合适的位置，如图 6.63 所示。添加后的效果如图 6.64 所示。单击“Edit Text”按钮可以修改文字内容，也可以设置文字的字体、字号、颜色、加粗、倾斜、对齐方式、添加超级链接、添加项目符号、添加多种阴影效果和设置字符间距等，如图 6.65 所示。单击“Animate”按钮可以设置文字动态效果，如图 6.66 所示。拖动文字的边框可以修改文字区域的长和宽。选择添加的文字，按【Delete】键可以删除该文字对象。

图 6.61　移动端效果

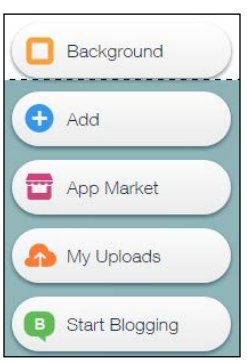


图 6.62　编辑按钮

图 6.63　选择添加的文字字体

图 6.64　添加文字后的效果

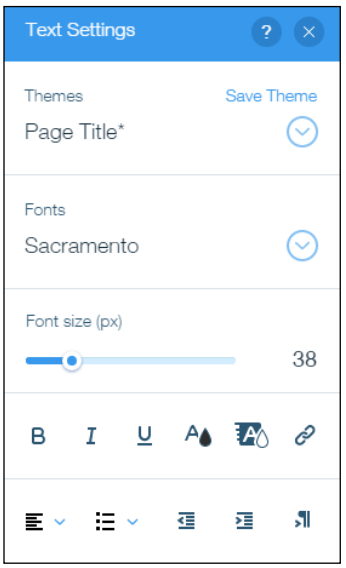


图 6.65　设置文字的字体、字号等

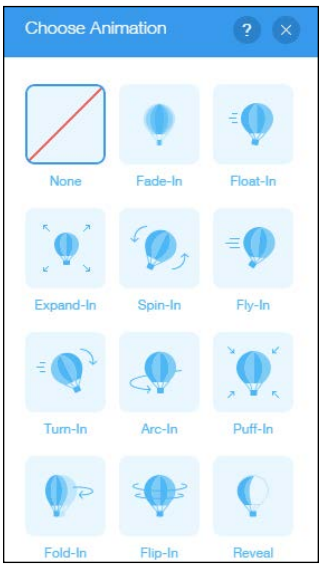


图 6.66　设置文字动态效果

网页上的任何对象都可以编辑和修改。网页修改后单击“Save”按钮进行保存，可以保存在 Wix 提供的免费空间，也可以连接到用户自己的网页空间，如图 6.67 所示。本案例保存到 Wix 提供的免费空间，并命名为“test1”，注意前面的 URL 是用户名“.wix.com”。单击“Save & Continue”按钮进行保存。

在保存网页的页面中，可以修改网页名称，也可以单击“Publish Now”按钮发布网页，然后单击“Done”按钮完成保存工作，如图 6.68 所示。若网页已经发布，则在浏览器中直接输入 URL 即可访问该页面，如本案例中的 URL 是“http://yinghliu.wix.com/test1”。如果不选择发布，则网页只能在用户登录账号后由本人浏览。

在网页中插入 Tableau 或 Echarts 图表。若有已经发布的嵌入代码（Embed Code）或者超级链接（Link），则可以将 Tableau 图表插入到 Wix 网页中。

在 5.10.2 小节中详细介绍了 Tableau 的发布过程。

图 6.67　选择网页的保存空间

图 6.68　保存网页

在 Wix 网页中单击“Add”下拉按钮，在打开的下拉菜单中选择“More”选项，再选择“Embed a Site”，如图 6.69 所示。将 Tableau 发布的图表的嵌入代码或 ECharts 发布的图表的嵌入代码复制到“Website Address”，如图 6.70 所示。最终的网页效果如图 6.71 所示。

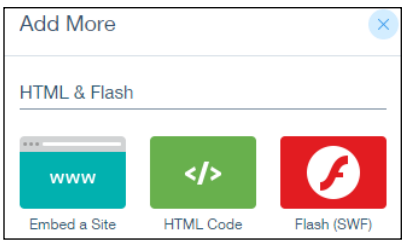


图 6.69　选择“Embed a Site”

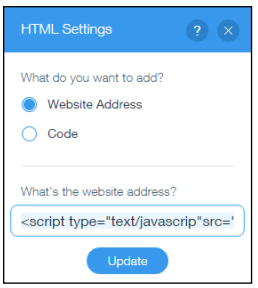


图 6.70　复制代码到“Website Address”

图 6.71　插入 Tableau 图表的网页